

PRIOR ART INFORMATION LIST

Your case No.	
Our case No.	20574

Inventor, Patent number, Country, Author, Title, Number of Document	Issue date	Concise Explanation of the Relevance (indication of page, column, line, figure of the relevant portion)
JP-A-8-153121 Atsuo KAWAI "AN AUTOMATIC DOCUMENT CLASSIFICATION METHOD BASED ON A SEMANTIC CATEGORY FREQUENCY ANALYSIS", Transactions of Information Processing Society of Japan (Information Processing Society of Japan) vol.33, No.9	June 11, 1996 Sep. 15, 1992	Abstract English translation of p.1114 Summary, 1(Instruction), and p.1119 3.3(Evaluation criteria) ~ p.1121 3.4(Evaluation of precision)/ Fig.5-1,5-2,and 6

(19)

JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **08153121 A**

(43) Date of publication of application: **11.06.96**

(51) Int. Cl.

G06F 17/30

G06F 12/00

(21) Application number: **07231033**

(22) Date of filing: **08.09.95**

(30) Priority: **30.09.94 JP 06236444**

(71) Applicant: **HITACHI LTD**

(72) Inventor:
MORITA TAKAKO
TONO JUNICHI
MATSUDA YOSHIKI
HASHIMOTO TETSUYA

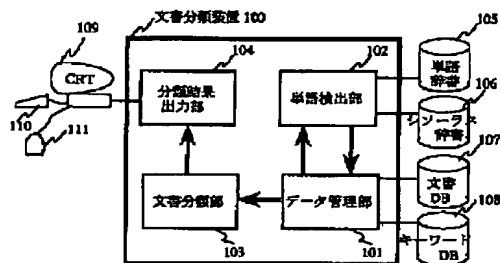
(54) **METHOD AND DEVICE FOR DOCUMENT
INFORMATION CLASSIFICATION**

(57) Abstract:

PURPOSE: To provide a method and device for document information classification which classify a document group without depending upon a prescribed classification system by using a key word given to the document group or a word appearing in a document and rearranges classification results hierarchically.

CONSTITUTION: A data management part 101 manages the document group in a document DB 107 and a group of key words, given to respective documents, in a key word DB 108. A document classification part 103 classifies the documents on the basis of the individual key words and stores them in folders. Then, folders having similar document groups are integrated. For the integration, it is judged whether the integration is effective or not. It is judged whether or not further classifications can be made in folders that are left without being integrated, thereby generating a hierarchical classification system. The classification results are outputted on a CRT 109 by a classification output part 104 to provide an environment wherein a user can read the classification results out.

COPYRIGHT: (C)1996,JPO



(19)日本国特許庁(JP)

(12) 公開特許公報(A)

(11)特許出願公開番号

特開平8-153121

(43)公開日 平成8年(1996)6月11日

(51)Int.Cl. [°]	識別記号	庁内整理番号	FI	技術表示箇所
G 0 6 F 17/30				
12/00	5 2 0 E	7623-5B		
		9194-5L	G 0 6 F 15/ 403	3 2 0 C
		9194-5L	15/ 401	3 1 0 D

審査請求 未請求 請求項の数25 O L (全 37 頁)

(21)出願番号 特願平7-231033

(22)出願日 平成7年(1995)9月8日

(31)優先権主張番号 特願平6-236444

(32)優先日 平6(1994)9月30日

(33)優先権主張国 日本(JP)

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72)発明者 森田 貴子

神奈川県川崎市麻生区王禅寺1099番地 株式会社日立製作所システム開発研究所内

(72)発明者 東野 純一

神奈川県川崎市麻生区王禅寺1099番地 株式会社日立製作所システム開発研究所内

(72)発明者 松田 芳樹

神奈川県川崎市麻生区王禅寺1099番地 株式会社日立製作所システム開発研究所内

(74)代理人 弁理士 小川 勝男

最終頁に続く

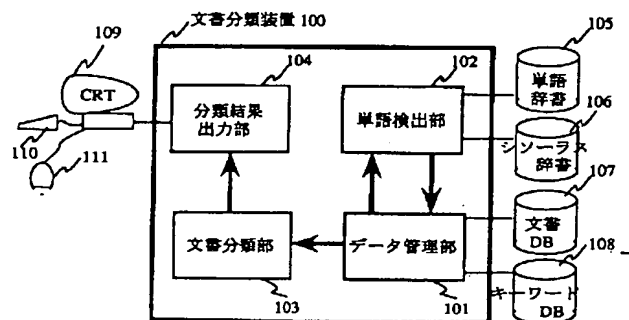
(54)【発明の名称】 文書情報分類方法および文書情報分類装置

(57)【要約】

【目的】 本発明の目的は、文書群に付与されたキーワードおよび文書中の出現単語を手がかりとして、規定の分類体系に依存せずに文書群进行分类し、分類結果を階層的に整理する文書情報分類方法および文書情報分類装置を提供することにある。

【構成】 文書DB107の文書群とキーワードDB108の各文書に付与されたキーワード群をデータ管理部101で管理する。文書分類部103では、個々のキーワードに基づいて文書をフォルダに分類して格納する。次に、類似した文書群を持つフォルダを統合する。統合する際には統合が有効かどうかの判断を行う。統合されたフォルダ内や、統合されずに残ったフォルダ内は、細分類が可能かどうかを判断し、階層的な分類体系を作り出す。分類結果は分類結果出力部104でCRT109に出力し、ユーザが分類結果を読み出すことができる環境を提供する。

図1



1

【特許請求の範囲】

【請求項1】 予め蓄積された複数の文書情報のそれぞれを所定の分類に分類する文書情報分類方法において、上記文書情報および上記文書情報の各文書情報におけるキーワード群を蓄積しておき、蓄積された各キーワード毎に、キーワードを含む上記文書情報を単一キーワードフォルダにまとめ、上記単一キーワードフォルダ内の文書同士を比較することにより、上記単一キーワードフォルダ同士を結合して関連キーワードフォルダを作成し、予め蓄積された上記複数の文書情報を上記関連キーワードフォルダに分類することを特徴とする文書情報分類方法。

【請求項2】 請求項1に記載の文書情報分類方法において、上記関連キーワードフォルダ同士を結合し、上記複数の文書情報を新たな関連キーワードフォルダに分類することを特徴とする文書情報分類方法。

【請求項3】 請求項2に記載の文書情報分類方法において、予め関連キーワードフォルダの数を指定しておき、上記関連キーワードフォルダ同士を結合する際、指定された関連キーワードフォルダの数になるまで結合を繰り返すことを特徴とする文書情報分類方法。

【請求項4】 請求項2に記載の文書情報分類方法において、上記新たな関連キーワードフォルダ内の各文書の内容により、上記新たな関連キーワードフォルダを存続させるか否か確認することを特徴とする文書情報分類方法。

【請求項5】 請求項1に記載の文書情報分類方法において、上記関連キーワードフォルダ内の文書情報同士を比較することにより、上記文書情報を第2の単一キーワードフォルダにまとめ、上記第2の単一キーワードフォルダ内の文書情報同士を比較することにより、上記第2の単一キーワードフォルダ同士を結合して、第2の関連キーワードフォルダを作成し、上記関連キーワードフォルダ内の文書情報を上記第2の関連キーワードフォルダに分類することを特徴とする文書情報分類方法。

【請求項6】 請求項5に記載の文書情報分類方法において、予め上記第2の関連キーワードフォルダ内に分類される文書情報の数を定めておき、上記第2の関連キーワードフォルダに分類するステップは、予め定められた数になるまで上記文書情報を分類することを特徴とする文書情報分類方法。

【請求項7】 請求項1に記載の文書情報分類方法において、

2

予め蓄積された上記複数の文書情報中に現れる言葉を抽出し、抽出された上記言葉を上記キーワードとすることを特徴とする文書情報分類方法。

【請求項8】 請求項1に記載の文書情報分類方法において、上記単一キーワードフォルダ内の文書同士を比較して、一致する文書情報の数が所定以上の単一キーワードフォルダ同士を結合して上記関連キーワードフォルダを作成することを特徴とする文書情報分類方法。

【請求項9】 請求項1に記載の文書情報分類方法において、上記キーワードの上記文書情報中での出現頻度および出現位置のうちいずれか一方を用いることにより上記関連キーワードフォルダを作成することを特徴とする文書情報分類方法。

【請求項10】 請求項1に記載の文書情報分類方法において、上記関連キーワードフォルダ内の各文書情報の特徴ベクトルを算出して、算出された各特徴ベクトルの平均ベクトルを求め、求められた上記平均ベクトルと特徴ベクトルとの差が所定以上の文書情報を上記関連キーワードフォルダ内で再分割することを特徴とする文書情報分類方法。

【請求項11】 請求項1に記載の文書情報分類方法において、上記関連キーワードフォルダ内の各文書情報の特徴ベクトルを算出し、算出した上記特徴ベクトルを用いて上記関連キーワードフォルダの結合の可否を判定することを特徴とする文書情報分類方法。

【請求項12】 予め複数の文書情報および上記文書情報の各文書情報におけるキーワード群を蓄積しておく記憶手段を有し、上記文書情報を分類する文書情報分類装置において、上記記憶手段に蓄積された各キーワード毎に、キーワードを含む上記文書情報を単一キーワードフォルダにまとめる単一キーワード処理手段と、上記単一キーワードフォルダ内の文書情報同士を比較することにより、上記単一キーワードフォルダ同士を結合して関連キーワードフォルダを作成する関連キーワード処理手段とを有し、予め蓄積された上記複数の文書情報を上記関連キーワードフォルダに分類することを特徴とする文書情報分類装置。

【請求項13】 請求項12に記載の文書情報分類装置において、上記関連キーワードフォルダ同士を結合する関連キーワード結合手段を有し、上記複数の文書情報を新たな関連キーワードフォルダに分類することを特徴とする文書情報

10

20

30

40

50

3

報分類装置。

【請求項14】請求項13に記載の文書情報分類装置において、

上記新たな関連キーワードフォルダ内の各文書情報の内容により、上記新たな関連キーワードフォルダを存続させるか否か確認する確認手段を有することを特徴とする文書情報分類装置。

【請求項15】請求項12に記載の文書情報分類装置において、

上記関連キーワードフォルダ内の文書情報同士を比較することにより、上記文書情報を第2の単一キーワードフォルダにまとめる第2の単一キーワード処理手段と、
上記第2の単一キーワードフォルダ内の文書情報同士を比較することにより、上記第2の単一キーワードフォルダ同士を結合して、第2の関連キーワードフォルダを作成する第2の関連キーワード処理手段とを有し、
上記関連キーワードフォルダ内の文書情報を上記第2の関連キーワードフォルダに分類することを特徴とする文書情報分類装置。

【請求項16】請求項12に記載の文書情報分類装置において、

予め蓄積された上記複数の文書情報中に現れる言葉を抽出する抽出手段と、
抽出された上記言葉を上記キーワードとするキーワード作成手段を有することを特徴とする文書情報分類装置。

【請求項17】請求項12に記載の文書情報分類装置において、

上記関連キーワードフォルダ処理手段は、上記単一キーワードフォルダ内の文書情報同士を比較して、一致する文書情報の数が所定以上の単一キーワードフォルダ同士を結合して上記関連キーワードフォルダを作成することを特徴とする文書情報分類装置。

【請求項18】請求項12に記載の文書情報分類装置において、

上記関連キーワードフォルダ処理手段は、上記キーワードの上記文書情報中での出現頻度および出現位置のうちいずれか一方を用いることにより上記関連キーワードフォルダを作成することを特徴とする文書情報分類装置。

【請求項19】請求項12に記載の文書情報分類装置において、

上記関連キーワードフォルダ内の各文書の特徴ベクトルを算出して、算出された各特徴ベクトルの平均ベクトルを求める平均ベクトル算出手段と、
求められた上記平均ベクトルと特徴ベクトルとの差が所定以上の文書情報を上記関連キーワードフォルダ内で再分割する再分割手段とを有することを特徴とする文書情報分類装置。

【請求項20】請求項12に記載の文書情報分類装置において、

上記関連キーワードフォルダ内の各文書の特徴ベクトル

4

を算出する特徴ベクトル算出手段と、

算出した上記特徴ベクトルを用いて上記関連キーワードフォルダの結合の可否を判定する結合判定手段とを有することを特徴とする文書情報分類装置。

【請求項21】予め複数の文書情報および上記文書情報の各文書情報におけるキーワード群を蓄積しておく記憶手段を有し、上記文書情報を分類する文書情報分類装置において、

上記記憶手段に蓄積された各キーワード毎に、キーワードを含む上記文書情報を単一キーワードフォルダにまとめる単一キーワード処理手段と、

上記単一キーワードフォルダ内の文書情報同士を比較することにより、上記単一キーワードフォルダ同士を結合して関連キーワードフォルダを作成する関連キーワード処理手段と、

操作者が指定した上記関連キーワードフォルダ同士を結合する手段を有し、

予め蓄積された上記複数の文書情報を上記関連キーワードフォルダに分類することを特徴とする文書情報分類装置。

【請求項22】請求項12に記載の文書情報分類装置において、

上記単一キーワード処理手段は、操作者が選択した文字列を含む文書情報を単一キーワードフォルダにまとめることを特徴とする文書情報分類装置。

【請求項23】請求項22に記載の文書情報分類装置において、

文字列の階層構造を示すシンソーラス辞書を有し、

上記シンソーラス辞書を用いて、上記関連キーワードフォルダの階層関係を構築する階層関係構築手段とを有することを特徴とする文書情報分類装置。

【請求項24】請求項22に記載の文書情報分類装置において、

上記関連キーワードフォルダに含まれる文書情報に基づいて関連キーワードフォルダ同士の類似度を定める手段と、

上記類似度に応じて、操作者が選択した関連キーワードフォルダに類似する関連キーワードフォルダを選択する手段を有することを特徴とする文書情報分類装置。

【請求項25】請求項24に記載の文書情報分類装置において、

上記類似度に応じて、操作者が選択した文書情報と同じ関連キーワードフォルダに属し、類似する文書情報を上記関連キーワードフォルダから取り除く手段とを有することを特徴とする文書情報分類装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、計算機上に蓄積された文書ファイルを階層的に分類する文書情報分類方法および文書情報分類装置に関する。特に、情報分類システム

5

および文書ファイリングシステム等に適用されるものである。

【0002】

【従来の技術】 計算機ネットワークの整備によって、情報検索サービスや電子メールを利用して、情報収集が行える環境が整いつつある。計算機ネットワークを介して、新しい情報が次々に到着し、話題は時々刻々変化する。最新情報の有用性は十分認識されているが、次の問題によって有効活用されていないのが現状である。

【0003】 (1) 所望の情報を選択するための検索式を論理式で入力しなければならない。このことは一般ユーザには困難である。

【0004】 (2) 所望する情報を得るために適切な検索式を作成することが困難である。

【0005】 (3) 収集した情報数が多くなると読み切れず、収集した中で必要な情報だけを選択して読むことができない。

【0006】 「21世紀の情報化社会」(日経バイト、1991年11月、320～331ページ)に記載されている広域情報サーバは、(1)、(2)の問題点を関連性フィードバックにより解決している。関連性フィードバックとは、ユーザが提示した検索条件に基づいて検索を行い、検索結果の中からユーザが所望の情報を選択すると、選択した情報を検索条件にフィードバックして、検索条件を改善するものである。はじめにユーザが提示した検索条件が不適切であっても、後に学習が行われて検索条件が洗練されるという手法である。

【0007】 「情報のブロードキャッチシステム」(情報処理学会 情報メディア研究会13-6、グループウェア研究会4-6報告、1993年10月28日、37ページ～44ページ)に記載されているシステムでは、ユーザが自分の興味をキーワードとしてシステムに登録しておく、これに適合する文書を収集する。論理式の入力を必要とせず、キーワードの登録だけでよいことから、(1)の問題点が解決できる。さらに、収集した文書数が一定数を超えると、文書群を自動分類している。文書の内容をキーワードの出現頻度の並びによるワードベクトルで表現し、ワードベクトル間の類似度を計算して類似する文書をまとめ、文書を分類する。

【0008】 また、特開平5-28198号公報記載の文書情報検索装置は、各文書に付与された分類や文書中に出現する単語といった検索情報を手がかりとして文書検索を行う。ユーザが検索したい分野や単語などの検索データを入力すると、内容を推測し自動的に検索式を作成して検索を行うことにより、(1)の問題を解決している。検索式の作成に際しては、あらかじめ登録してある検索論理式のうち、最適な論理式を選択する。さらに、検索された文書群は検索データとの類似度を算出し、その値を使って整列することで、(3)の問題に対処している。

6

【0009】 また、これらの従来例では、1つの視点でしか文書分類できない。そのため、検索漏れを起こすとの問題もある。

【0010】

【発明が解決しようとする課題】 前述の「21世紀の情報化社会」に記載された広域情報サーバでは、関連性フィードバックという方法によって、ユーザが指定した検索条件だけではなく、ユーザの希望にかなった実情報を利用して、システムが検索条件を改善し、(1)、

(2)の問題を解決した。

【0011】 「情報のブロードキャッチシステム」では、検索式を入力するのではなく、キーワードを登録するという方法を採用している。しかし、ユーザが自分の興味を的確に反映するキーワードを登録するのは困難なため、(2)の問題は解決し切れていない。また、文書群の自動分類処理は逐次的に起動されることが前提なので、文書の到着順序に依存した分類になる。情報の内容は時々刻々変化するもので、過去のある時点における分類体系に依存し続けると、有効な分類結果は生成できず、(3)の解決は困難になる。

【0012】 上述したように、文書の収集、検索処理ではユーザが自分の興味を的確に反映したキーワードを設定する作業が困難であるという問題がある。

【0013】 この検索処理に対する問題と比較して、文書の分類処理でも分類の手がかりとなる分類体系をあらかじめ設定する作業においても同様の問題が生じる。つまり、一つの分類体系に依存し続けると、内容の変化に対応することができず、不適切な分類結果になり得るという問題が生じる。

【0014】 一方、特開平5-28198号公報記載の文書情報検索装置では、検索結果を有効なものから順に整列して検索式に類似する文書から見られるように工夫している。しかし、何番目までが有効な情報かを判断するには、ユーザは文書の内容を確認せざるを得ないため、(3)の問題を解決できるとはいえない。

【0015】 つまり、検索結果を一階層に整列しただけでは、類似した内容の文書をまとめて見たり、まとめて読み飛ばしたりすることができないという問題がある。

【0016】 これらの問題を解決するものとして、特開平5-324726号公報がある。この従来例では、あるキーワードに着目して文書内にそのキーワードが存在するか否かにより文書を分類する。分類された文書に同様の処理を施すことにより階層的に分類していく。

【0017】 しかし、この従来例でも複数分野に関連する文書を1つの視点でしか分類できないとの問題点を有する。

【0018】 また、本従来例を新聞などの一般の文書データに適用すると階層が深くなり、分類数が爆発的に増加するとの問題もある。

【0019】 本発明の目的は上述した問題を解決するた

7

めに、既定の分類体系にとらわれることなく文書を自動的に分類し、分類結果を階層的に整理する文書分類方法および文書分類装置を提供することにある。

【0020】

【課題を解決するための手段】上記の目的を達成するために本発明は、予め複数の文書および文書の各文書におけるキーワード群を蓄積しておき、蓄積された各キーワード毎に、キーワードを含む文書を単一キーワードフォルダにまとめ、単一キーワードフォルダ内の文書同士を比較することにより、単一キーワードフォルダ同士を結合して関連キーワードフォルダを作成し、予め蓄積された複数の文書を関連キーワードフォルダに分類することを特徴とする。

【0021】本発明の文書情報分類方法および文書情報分類装置では、蓄積された複数の文書と、各文書に人手で付与したキーワードと各文書中から自動抽出した単語をまとめたキーワード群と、キーワード群中の個々のキーワードから算出した重要度を管理し、キーワード群と重要度を分類処理に利用する。

【0022】分類処理では、まず各キーワードごとに文書をまとめる単一キーワード分類処理によって、文書群を単一キーワードフォルダに格納する。複数のキーワードを持つ文書は、複数の単一キーワードフォルダに重複して分類する。

【0023】次に、類似した文書群を含む単一キーワードフォルダについて統合判定を行い、統合可能と判断した場合には関連キーワード分類処理で統合し、関連キーワードフォルダに格納する。

【0024】さらに、類似した文書群を含む関連キーワードフォルダについて統合判定を行い、統合可能と判断した場合には関連キーワード分類処理手段で関連キーワードフォルダの統合を繰り返す。

【0025】単一キーワードフォルダ内あるいは関連キーワードフォルダ内について、細分類判定を行い、細分類可能な場合は単一キーワード分類処理と関連キーワード分類処理を利用して、階層的に分類する。細分類は分類停止の判断するまで再帰的に繰り返す。細分類不可能な場合は文書間関連度の判定を行い、関連度が低い文書を雑音とみなして分ける。

【0026】各関連キーワードフォルダには、フォルダ内に格納された文書群を代表する名称を付与し、フォルダ名を付ける。

【0027】

【作用】最終的に、既定の分類体系に依存することなく、各文書を必ず一つ以上の分類に格納し、階層的な分類体系を作成し、分類結果群を代表する名称を付与することができる。その結果、ユーザが大量の文書から所望の文書を見つけやすくなることができる。

【0028】

【実施例】以下本発明の一実施例について説明する。

8

【0029】本実施例の文書分類装置が対象とするのは、計算機上のテキストファイルであり、以下、これを文書とする。各文書には、文書の内容を代表する複数のキーワードを付与することができ、以下、これをキーワード群とする。

【0030】図1に本実施例の文書分類装置の構成例を示す。文書分類装置100は、データ管理部101、単語検出部102、文書分類部103、分類結果出力部104から構成されており、一般用語を収録した単語辞書105、用語間の上位下位関係や同義語情報などを収録したシソーラス辞書106、文書を格納している文書DB107、各文書のキーワード群を格納しているキーワードDB108、出力装置のCRT109、入力装置のキーボード110、マウス111を持つ。

【0031】データ管理部101は、文書DB107と、キーワードDB108を管理し、文書やキーワード群の入出力を行う。キーワードDB108には、あらかじめ人手で付与したキーワード（以下、人手付与キーワードとする）を格納することができる。人手付与キーワードは格納する必要はないが、本実施例では格納した場合を例に説明する。

【0032】単語検出部102は、データ管理部101から文書を受け取り、単語辞書105を利用して形態素解析処理を行い、文書から単語を抽出する。または、文書を文字種の変化点で分割し、単語を抽出する手法や、文書中のn文字の並びに対する出現頻度を調べ、意味を持つn文字の並びを抽出する手法などを利用してよい。抽出した単語群は、シソーラス辞書106を参照して、同義語展開したり、上位語への変換を行い、同義語群を生成する。抽出した単語群と、同義語群をあわせて、単語キーワードを生成する。単語キーワードをデータ管理部101に受け渡し、キーワードDB108中の各文書のキーワード群に追加登録する。

【0033】また、単語検出部102は各単語キーワードの重要度を計算する。重要度とは、（1）全単語の出現頻度総数における各単語の出現頻度、（2）タイトル、見出し、段落などの文書中での出現位置に基づく。すなわち、タイトルは文書の内容を代表する可能性が高いので、タイトル中に出現した単語の重要度を高くするといった手法である。

【0034】文書分類部103は、データ管理部101から各文書のキーワード群を受け取り、階層的な分類体系を生成し、結果を分類結果出力部104に受け渡す。分類処理の詳細は後述する。

【0035】分類結果出力部104は、文書分類部103から階層的な分類体系を受け取り、インタフェース画面をCRT109に表示する。

【0036】図2に単語キーワードの具体例を示し、図1の単語検出部102の詳細を説明する。まず、単語辞書105を参照して文書201から単語群202の抽出

を行う。単語群202は、シソーラス辞書106を参照して同義語や上位語に展開する。同義語群203は、

「ブレンド」という用語が同義語の「調合」に展開された例を示していて、同義語への展開はキーワードの表記を統一するために行うものである。次に、単語群202と同義語群203から、単語キーワード204を生成し、これをキーワードDB108に格納されている文書201のキーワード群205に格納する。キーワード群205は、あらかじめ付与されていた人手付与キーワードに、単語キーワード204が追加された様子を示す。

【0037】図3の流れ図に従い、文書分類部103の詳細について説明する。分類の結果、文書が格納される領域をフォルダ、フォルダに格納される複数の文書を文書群と呼ぶ。また、「フォルダを作成する」とは、分類された文書群を格納する領域を確保することである。

【0038】まず、各キーワードをキーワード群中に含む文書を、各キーワードごとにまとめる単一キーワード分類処理301を行う。次に、単一キーワード分類処理301によって作成された単一キーワードフォルダについて、類似した文書群を含む単一キーワードフォルダの組を統合する関連キーワード分類処理302を行う。ステップ303では、関連キーワード分類処理302によって作成された関連キーワードフォルダについて、類似した文書群を含む関連キーワードフォルダの組を統合できるかどうかの判定を行い、統合が可能な間は関連キーワードフォルダの統合を繰り返す。さらに、作成された単一キーワードフォルダや関連キーワードフォルダ内について、細分類できるかどうかを調べ（ステップ304）、細分類可能な間は階層的に分類を繰り返す（ステップ305）。ステップ305は、すべてのフォルダ内について階層的な分類を行う。

【0039】次に、図3の各処理の詳細を説明する。

【0040】まず、図4の流れ図を用いて、図3の単一キーワード分類処理301の詳細を説明する。以下、単一キーワード分類処理301で作成するフォルダを単一キーワードフォルダと呼ぶ。

【0041】ステップ401では、キーワードを管理するテーブルの初期化を行う。キーワードを管理するテーブルをキーワードテーブルと呼ぶ。ステップ402では、図1のキーワードDB108から一文書のキーワード群を読み出し、キーワードテーブルに各キーワードを登録する。すべての文書について、キーワードの登録を繰り返す（ステップ403）。ステップ404では、各キーワードについて、キーワードをキーワード群に含む文書数を数える。文書数が1であるかどうかの判定を行い（ステップ405）、1のときはキーワードを対象外とする（ステップ406）。一文書にしか含まれないキーワードは、分類時に他の文書との関連性を判断する材料にならないためである。文書数が1ではないときは、キーワードと同一の名前の単一キーワードフォルダを作成

し、キーワードをキーワード群中に含む文書群を単一キーワードフォルダに格納し、さらに単一キーワードフォルダ名をフォルダを管理するテーブルに登録する（ステップ407）。フォルダを管理するテーブルをフォルダテーブルと呼ぶ。ステップ408では、全キーワードについて、上述した単一キーワードフォルダの作成処理を繰り返す。

【0042】次に、上述したキーワードテーブルについて図5を用いて説明する。キーワードテーブル501の各エントリは、キーワードリスト502を指している。エントリは、キーワードを入力とするハッシュ関数の値で決定する。キーワードリスト502は、キーワード503、同一ハッシュ値のキーワードリストへのポインタ504、文書識別番号リストへのポインタ505の組である。ポインタ504は、同一ハッシュ値のキーワードリスト506を指していて、キーワードリスト502と同一ハッシュ値のキーワードリスト506のキーワードは同じハッシュ値になることを示す。ポインタ505は、文書識別番号リスト507を指していて、キーワード503を含む文書群が連結されている。文書識別番号リスト507は、文書を識別するための番号508、次文書識別番号リストへのポインタ509の組である。各キーワードを含む文書の数は、キーワードテーブル501の文書識別番号リスト505をたどれば得られる。

【0043】図6に示したキーワードテーブルの具体例では、キーワードテーブル600の7番のエントリ601にキーワードリストが連結されている。ポインタ603は、「ペット」というキーワード602のキーワードリストに、「新種」というキーワード609のキーワードリストが連結していることを示し、ポインタ610はキーワードリストの連結はそれ以上はないことを示している。ポインタ604は、「ペット」というキーワードを含む一つの文書識別番号リストである文書識別番号1（605）を指し、さらにポインタ606は文書識別番号2（607）を指している。ポインタ608は、それ以上「ペット」というキーワードを含む文書群はないことを示す。同様に、ポインタ611は「新種」というキーワードを含む文書識別番号リストの文書識別番号3（612）を指している。ポインタ613は、さらに文書識別番号リストが連結されていることを示す。

【0044】次に、フォルダテーブルについて図7を用いて説明する。フォルダテーブル700は、フォルダ名701、フォルダに格納されている文書数702、分類終了フラグ703、文書識別番号リスト704の組である。分類終了フラグ703は分類を行うか否かを記憶するフラグで、詳細については後述する。文書識別番号リスト704は、図5のキーワードテーブル501の文書識別番号リスト507と等しい。フォルダテーブル700の0番のエントリ705は、「ペット」というフォルダがあり、その中には文書が2件格納されていて、フォ

1 1

ルダの分類はOFF（終了していない）で、文書識別番号のリストが連結していることを示す。

【0045】図8に示す単一キーワード分類処理の具体例を用いて、図3の単一キーワード分類処理301を説明する。文書801は図1の文書DB107に格納されていて、文書群802を形成する。各文書にはキーワード群803が付与されていて、キーワード群803は個々のキーワード804の集合である。8031は文書1のキーワード群で、「犬」、「動物」、「ペット」というキーワードを含む。8032は文書2のキーワード群で、「猫」、「動物」、「ペット」、「ねずみ」というキーワードを含む。8033は文書3のキーワード群で、「新種」、「ねずみ」、「動物」、「ABC国」というキーワードを含む。8034は文書4のキーワード群で、「ねずみ」、「キャラクター商品」、「ABC国」というキーワードを含む。その他の文書5、文書6、文書7、文書8も同様にキーワード群を持つものとする。文書群802に対して単一キーワード分類処理を行うと、単一キーワードフォルダ800群に展開される。単一キーワードフォルダは、「犬」、「猫」、「キャラクター商品」、「新種」、「ABC国」、「ペット」、「ねずみ」、「動物」、というキーワードと同一の名前で作成されている。各単一キーワードフォルダには、文書集合805が格納されている。

【0046】例えば、8051は単一キーワードフォルダ「犬」の文書集合で、文書1が含まれていることを示す。8052は、単一キーワードフォルダ「ペット」の文書集合で、文書1と文書2が格納されている。8053は単一キーワードフォルダ「ねずみ」の文書集合で、文書2、文書3、文書4が格納されている。8054は単一キーワードフォルダ「動物」の文書集合で、文書1、文書2、文書3が格納されている。8055は単一キーワードフォルダ「ABC国」の文書集合で、文書3、文書4が格納されている。8056は単一キーワードフォルダ「新種」の文書集合で、文書3、文書5、文書6、文書7、文書8が格納されていることを示す。

【0047】次に、図9の流れ図に従い、図3の関連キーワード分類処理302の詳細を説明する。以下、関連キーワード分類処理302で作成するフォルダを関連キーワードフォルダと呼ぶ。

【0048】ステップ901では、フォルダ間の一致文書数を管理するテーブルの初期化を行う。フォルダ間の一致文書数を管理するテーブルをフォルダ間一致文書数管理テーブルと呼ぶ。二つの単一キーワードフォルダ間での一致文書数を数えて、フォルダ間一致文書数管理テーブルに登録する（ステップ902）。ステップ903では、すべての単一キーワードフォルダ間の一致文書数をフォルダ間一致文書数管理テーブルに登録する。ステップ904では、フォルダ間一致文書数管理テーブルを一致文書数で降順に配列し、テーブルの先頭すなわちフ

1 2

ォルダ間の一致文書数が最大となった単一キーワードフォルダの組の統合が可能であるかを調べる（ステップ905）。ステップ905の詳細は後述する。

【0049】統合が可能である場合は、関連キーワードフォルダを作成し、統合する二つの単一キーワードフォルダの文書群を格納し、関連キーワードフォルダ名を図7のフォルダテーブル700に登録し、統合した二つの単一キーワードフォルダをフォルダテーブル700から削除する（ステップ907）。関連キーワードフォルダの名称は、統合した二つの単一キーワードフォルダ名を列挙したものである。このとき、列挙する順番は文書数の多い単一キーワードフォルダ名から並べ、関連キーワードフォルダ内の文書群がどのようなキーワードを多く含むかを明示する。さらに、ステップ908でフォルダ間一致文書数管理テーブルに統合が終了したことを書き込み、フォルダ間一致文書数管理テーブルの更新を行う。

【0050】統合が不可能である場合や、統合が終了した場合には、統合処理を行っていない単一キーワードフォルダの組について、フォルダ間一致文書数管理テーブルの先頭から終りまで、統合処理を繰り返す（ステップ906）。

【0051】上述したフォルダ間一致文書数管理テーブルを図10に示し、詳細を説明する。フォルダ間一致文書数管理テーブル1001は、一致文書数を調べる二つのフォルダの識別番号1002、1003、一致文書数1004、統合済フラグ1005の組である。図10の例では、フォルダ識別番号0番と1番の一致文書数が5件あったことを示している。統合済フラグ1005は、「フォルダの統合が済んでいるか否か」や「フォルダの統合が不可能である」ことを記憶するフラグで、「済」は統合が済んでいることを示し、「不可」は統合が不可能であることを示している。

【0052】次に、図11にフォルダ間一致文書数管理テーブルの具体例を示す。すべての単一キーワードフォルダの識別番号を1101と識別番号1102に列挙し、二つのフォルダ間の一致文書数を調べて登録したもので、例えば、文書0と文書4の一致文書数1103は8件あったことを示す。一致文書数は、図7のフォルダテーブル700の文書識別番号リスト704をたどり、比較することで求められる。

【0053】図12にフォルダ間一致文書数管理テーブルの更新処理の具体例を示し、これを用いて詳細を説明する。フォルダ間一致文書数管理テーブル1200は、図9のステップ904まで処理が終了した状態を示す。フォルダ間一致文書数管理テーブルの先頭、すなわちエントリ0番からフォルダの統合処理を開始し、フォルダ識別番号0番と1番の統合が可能と判断されると統合を行い（ステップ905）、統合済フラグ1205に「済」と書き込む（ステップ908）。フォルダ識別番

13

号0番と1番を統合すると、0番と1番に関係する1206~1213を「不可」にする。次に、フォルダ間一致文書数テーブルの次のエントリ、すなわちエントリ1番について統合処理を行う(ステップ906)。このとき、統合済フラグが「不可」ではないことを確認する。

「不可」ならばエントリ1番の単一キーワードフォルダの組の一方は、すでに統合処理が終了しているので統合はできない。図12では、統合済フラグ1214は「不可」ではないので、統合を行うことができる。本実施例では、フォルダ間一致文書数管理テーブルの先頭から順番に処理を行うので、「済」のエントリが処理中のエントリよりも後ろのエントリに現われることはない。

【0054】次に、図13の流れ図を用いて、図9のステップ905のフォルダの統合判定処理の詳細を説明する。

【0055】統合するフォルダの組は、「一致文書数が最大になること」を条件として決定する。統合すべきかどうかは、統合した結果が有効な分類になるかどうかを調べる必要があり、統合前と統合後を比較して判定する。本実施例では図13に示すフォルダ内文書の距離計算処理1300を適用する。

【0056】図13の1301は、フォルダ内文書の距離計算処理1300で一時的に用いる作業用キーワードテーブルの初期化を行っている。作業用テーブルのデータ構造は、図5のキーワードテーブル501と同様とする。フォルダ内に格納されている文書に対して、キーワードの登録を行い(ステップ1302)、すべてのフォルダ内文書について繰り返す(ステップ1303)。キーワード数の計数用にキーワード数pを初期化する(ステップ1304)。キーワードを含む文書数を数えて(ステップ1305)、文書数が1のキーワードは作業用キーワードテーブルから削除し(ステップ1308)、1以上のときはキーワード数pを1ずつ増やす(ステップ1307)。すべてのキーワードについてステップ1305以下の処理を繰り返し(ステップ1309)、フォルダ内の文書群に含まれるキーワードのうち、二つ以上の文書に含まれるキーワードの選定が終了する。次に、図14のステップ1400に進む。

【0057】図14の1401では、ワードベクトルを管理するテーブルの初期化を行う。ワードベクトル W_i とは、文書 i における「キーワードの出現頻度とキーワードの重要度の積」の並びであり、具体的には次のように表現できる。

【0058】ワードベクトル $W_i = (F1*V1, F2*V2, \dots, Fj*Vj, \dots, Fp*Vp)$

(i は文書識別番号、 $1 \leq j \leq p$, j はキーワード識別番号、 p はキーワード数、 F_j はキーワード j の出現頻度、 V_j はキーワード j の重要度)

重要度は、図1の単語検出部102で付与したもので、値が大きいものほど重要度が高い。ワードベクトルを管

14

理するテーブルをワードベクトルテーブルと呼ぶ。ステップ1402では、各キーワードの出現頻度と単語検出部102で付与した各キーワードの重要度の積を計算し、ワードベクトルテーブルに登録し、各文書に関して繰り返す(ステップ1403)。さらに、ステップ1404では、各文書のワードベクトルの平均ベクトルを求める。本実施例では次式で定義する。

【0059】平均ベクトル $W_a = \sum W_i / n$

($1 \leq i \leq n$, i は文書識別番号、 n は文書数)

次に、ステップ1405では、各文書のワードベクトル W_i ($1 \leq i \leq \text{文書数}$)と平均ベクトル W_a との距離を計算する。ベクトル間の距離とは、ベクトルの近さを判断するもので、本実施例ではベクトル間の距離を次式で定義する。文書 D_i , D_j のワードベクトルをそれぞれ W_i , W_j とし、 W_i と W_j のなす角度を θ とし、 D_i , D_j 間の距離を $d(D_i, D_j)$ とする。

【0060】

$$d(D_i, D_j) = 1 - W_i \cdot W_j / |W_i| * |W_j| = 1 - \cos \theta$$

(ただし、 \cdot は内積、 $*$ は積、 $|W_i|$ は W_i の大きさ)

$d(D_i, D_j)$ は $0 \leq d(D_i, D_j) \leq 1$ の範囲の値であり、ベクトル間の距離が近いほど値は小さくなり、一致する場合は0になる。

【0061】すべての文書について、ワードベクトルと平均ベクトルとの距離計算を繰り返す(ステップ1406)。次にステップ1407で、すべての平均ベクトルと各文書との距離から距離分布を求める。距離分布とは、(1)平均距離、(2)分散とし、次式で定義する。

【0062】平均距離 $d_a = \sum d_i / n$

分散 $\sigma = \sum ((d_i - d_a) * (d_i - d_a)) / (n-1)$
($1 \leq i \leq n$, i は文書識別番号、 n は文書数、 d_i は文書識別番号 i の文書と平均ベクトルとの距離)

統合前の二つの単一キーワードフォルダについて、別々に調べた距離分布の平均値と、統合後の関連キーワードフォルダの距離分布を比較することで、統合の可否を判断する。本実施例では、(1)、(2)を具体的に次式で定義する。

【0063】(1) $|d_2 - d_1| \geq T_d$

(d_1 は統合前の二つの単一キーワードフォルダの平均距離の平均値、 d_2 統合後の平均距離、 $|x|$ は x の絶対値、 T_d はしきい値)

(2) $\sigma_2 / \sigma_1 \geq T_\sigma$

(σ_1 は統合前の二つの単一キーワードフォルダの平均分散値、 σ_2 は統合後の分散値、 T_σ はしきい値)

(1)または(2)の条件が満たされるときに、統合不可と判断する。しきい値 T_d , T_σ は、初期実験で数種類を決定しておき、フォルダ内の文書群の距離分布の状況に適したものを選択する。

【0064】図15のワードベクトルテーブルの具体例を用いて、詳細を説明する。ワードベクトルテーブル1

15

500は、縦軸が文書識別番号、横軸がキーワード識別番号の2次元テーブルである。例えばエントリ1503は、文書識別番号3番の文書において、キーワード識別番号3のキーワードの出現頻度と重要度の積が2であることを表している。図15を例にとり、ワードベクトルと平均ベクトルとの距離の計算例を示す。説明を簡略化するために、各キーワードの重要度はすべて1とし、文書数は4、キーワードは識別番号4までを対象としたときの、文書識別番号1番のワードベクトルと平均ベクトルとの距離を計算する。

【0065】 $W1 = (3, 2, 1, 1)$

$$d1(D1, Wa) = 1 - W1 \cdot Wa / |W1| \cdot |Wa|$$

$$= 1 - (3, 2, 1, 1) \cdot (1.3, 4.2, 3.5, 0.8) / 3.9 \cdot 3.8$$

$$= 1 - 48.1 / 148.2$$

$$= 1 - 0.32$$

$$= 0.68$$

図16に関連キーワード分類処理の具体例を示す。単一キーワードフォルダ「ペット」1601と単一キーワードフォルダ「動物」1602とを統合して関連キーワードフォルダ「動物とペット」1605を作り、単一キーワードフォルダ「ABC国」1603と単一キーワードフォルダ「ねずみ」1604を統合して関連キーワードフォルダ「ねずみとABC国」を作成している（図3ステップ302）。さらに、関連キーワード分類処理を繰り返す（図3ステップ303）、関連キーワードフォルダ「動物とペット」1605と関連キーワードフォルダ「ねずみとABC国」1606を統合し、関連キーワードフォルダ「動物とねずみとペットとABC国」1607を作成している。1607のフォルダ名は、「動物」、

「ねずみ」、「ペット」、「ABC国」の順に、各キーワードに関連する文書が多く格納されていることを示す。

【0066】関連キーワード分類処理が終了すると、分類体系の第一階層が生成される。

【0067】次に、細分類処理について説明する。細分類とはフォルダ内を階層的に分類することで、上位フォルダの作成に利用していないキーワードを用いて分類を行う。例えば、図16の関連キーワードフォルダ「動物とねずみとペットとABC国」1607内を細分類する場合には、キーワード「動物」、「ねずみ」、「ペット」、「ABC国」以外のキーワードを利用して分類を行う。

【0068】図17に細分類の流れ図を示し、図3のステップ304の詳細を説明する。ステップ1701ではフォルダの種類を判別する。

【0069】関連キーワードフォルダの場合は、類似度の高い、複数の単一キーワードフォルダが統合された結果なので、フォルダ内をさらに階層的に分類する。分類は、図3の流れ図に従って、単一キーワード分類処理301、関連キーワード分類処理302、関連キーワードフォルダの統合処理303、細分類304～306を再

16

$$*W2 = (1, 13, 2, 0)$$

$$W3 = (1, 1, 8, 0)$$

$$W4 = (0, 1, 3, 2)$$

$$\text{平均ベクトル } Wa = \sum W_i / 4$$

$$= (5, 17, 14, 3) / 4$$

$$= (1.3, 4.3, 3.5, 0.8)$$

(少数第2位を四捨五入)

文書識別番号1のワードベクトルと平均ベクトル Wa との距離 $d1$

10

*

※帰的に繰り返す。

【0070】単一キーワードフォルダの場合は、一つのキーワードに引き付けられた文書群が格納されているので、集合としてのまとまりがあるかどうかの保証がない。そこで、図13のフォルダ内の文書間の距離計算を行って（ステップ1300）、各文書と平均ベクトルとの距離の分散値を求める（ステップ1702）。

【0071】分散値としきい値 $T\sigma$ を比較し（ステップ1702）、 $T\sigma$ 以上ならばフォルダ内を階層的に分類する価値はないとみなし、さらに平均距離を調べる（ステップ1703）。平均距離がしきい値 Td 以上の文書は、関連性の薄い雑音文書であると判定し、単一キーワード内の雑音フォルダに格納する（ステップ1704）。雑音フォルダは、雑音と判定された文書を格納するためのフォルダで、雑音文書が存在したフォルダ内にだけ作成する。フォルダ内のすべての文書について、平均距離計算を行い（ステップ1705）、雑音文書を雑音フォルダに格納し、図7のフォルダテーブル700の分類終了フラグ703に分類終了と書き込む（ステップ1706）。

【0072】分散値が $T\sigma$ 以下の場合は、単一キーワードフォルダ内を細分類可能と判断し、図3の流れ図に従って、単一キーワード分類処理301、関連キーワード分類処理302、関連キーワードフォルダの統合処理303、細分類304～306を再帰的に繰り返す。

【0073】図18の細分類の具体例では、関連キーワードフォルダ「猫と魚」1801は、単一キーワードフォルダ「キャットフード」1804と、関連キーワードフォルダ「釣り」と海」1805の二つのフォルダに階層的に分類された例である。また、単一キーワードフォルダ「犬」1802は、単一キーワードフォルダ「柴犬」1806と関連キーワードフォルダ「えさと散歩」1807の二つのフォルダに階層的に分類された例である。単一キーワードフォルダ「新種」1803は、雑音文書

40

50

17

1809が雑音文書フォルダ1808に分けられた例である。

【0074】細分類によって、分類体系の第二階層以下を作成したことになる。

【0075】これまでは文書分類部の流れを説明した。ここで、分類によって作成されるフォルダの階層構造の記憶方法について、図19の分類階層管理テーブル1900を用いて説明する。

【0076】分類階層管理テーブル1900の各エントリは、フォルダ情報リストを指している。フォルダ情報リスト1901は、フォルダ名1902、文書識別番号リスト1903、文書数1904、兄弟フォルダ情報リストへのポインタ1905、子フォルダ情報リストへのポインタ1906、親フォルダ情報リストへのポインタ1907の組である。フォルダ名1902は図7のフォルダテーブル700のフォルダ名701と一致し、文書数1904は702と一致する。文書識別番号リスト1903は、各フォルダに格納されている文書識別番号リスト1908へのポインタであり、1908は文書識別番号1909と次文書識別番号リストへのポインタ1910の組である。文書識別番号リスト1903は、図7のフォルダテーブル700の文書識別番号リスト704と一致する。兄弟フォルダ情報リストへのポインタ1905は、フォルダ情報リスト1901と同じ上位フォルダを持ち、同階層に位置するフォルダ情報リストへのポインタである。子フォルダ情報リストへのポインタ1906は、フォルダ情報リスト1901の一つの下位フォルダ情報リストへのポインタである。親フォルダ情報リストへのポインタ1907は、上位フォルダ情報リストへのポインタである。

【0077】図7のフォルダテーブル700に作成したフォルダの情報を書き込むと同時に、分類階層管理テーブル1900にもフォルダテーブル700の内容を複写する。細分類によって、第二階層以降のフォルダを作成すると、分類階層管理テーブル1900の子フォルダ情報リストへのポインタ1906、兄弟フォルダ情報リストへのポインタ1905、親フォルダ情報リストへのポインタ1907の更新を行う。

【0078】図19の分類階層管理テーブル1900を用いて、具体的に階層構造の記憶方式を説明する。まず、新規作成したフォルダをフォルダ情報リスト1901に登録する。さらに、フォルダ内を細分類して二つのフォルダに分類したとすると、子フォルダ情報リストへのポインタ1906に一つの子フォルダ情報リスト1911を登録し、1911の兄弟フォルダ情報リストへのポインタ1912に二つ目の子フォルダ情報リスト1915を登録し、それ以上はフォルダはないので、1916は連結がないことを示す。フォルダ情報リスト1901は1911と1915の親フォルダ情報リストに相当するので、親フォルダ情報リストへのポインタ191

18

4、1918は1901を指している。フォルダ情報リスト1901の子フォルダ情報リスト1911、1915は、以下に階層的な分類はないので、1913、1918は連結がないことを示す。フォルダ情報リスト1901は第一階層のフォルダで、それ以上の階層や同一の階層に位置するフォルダはないので、1905、1907は連結がないことを示す。また、フォルダ情報リスト1901は、二つの文書を持ち、1903は一つ目の文書識別番号リスト1908を、1909は二つ目の文書識別番号リスト1910を指している。

【0079】図1の文書分類部103で生成された分類体系は、具体的には図20のような階層構造に展開され、第一階層には、関連キーワードフォルダ「猫と魚」1801、単一キーワードフォルダ「犬」1802、単一キーワードフォルダ「新種」が、第二階層フォルダには、1801の下位に単一キーワードフォルダ「キャットフード」1804、関連キーワードフォルダ「釣りと海」1805があり、1802の下位に単一キーワードフォルダ「柴犬」1806、関連キーワードフォルダ「えさと散歩」1807があり、1803の下位には雑音文書が雑音文書フォルダに分離されている。

【0080】図1の分類結果出力部104は、上記分類体系を文書分類部103から受け取ると、インタフェース画面を図21に示すように作成し、CRT109に出力する。図21の2101は、分類体系の上位三階層が表示されていて、2113は第一階層、2114は第二階層2115が第三階層を示していて、各階層のフォルダ名が縦方向に表示されている。図21は、第一階層

「猫、魚」を選択し、第二階層「キャットフード」を選択した結果、第三階層に文書群が表示され、文書15を選択した様子である。2112は文書15の内容、文書の作成された日2116やフォルダ内における文書15の得点情報2117を表示している。ユーザは図1のマウス111で、興味のあるフォルダ名を選択し、自由に内容を参照することができる。また、興味のないフォルダは読み飛ばすことができ、参照すべき文書量が削減できる。

【0081】図1の文書DB107に新規文書の到着や、古い文書の削除が行われる場合には、文書分類部103が改めて分類しなおすことで、新規情報の入手にも対応できる。

【0082】次に、第2の実施例を説明する。第2の実施例は、図1に示した第1の実施例の文書分類装置100における分類結果出力部104に、ユーザの意見を反映した分類結果を構築するための分類指定手段を設けた文書分類装置に関する。

【0083】第1の実施例の文書分類装置は既定の分類体系にとらわれることなく文書を自動的に分類するために、ユーザの意向と異なる分類結果を生成してしまうことがある。そこで図22のように、図1の分類結果出力

1.9

部104に分類指定部2201を加えた、文書分類装置2200の構成をとり、ユーザの意見を分類結果に反映させる手段を設ける。分類指定部2201はフォルダ数指定インタフェース2500と分類体系構築補助インタフェース2700という二つの画面から構成される。

【0084】ユーザがキーボード110、マウス111といった入力装置を用いて分類したい文書群を指定し、分類処理の実行を指示すると、文書分類装置2200が起動され、図23の流れ図に基づく処理が施される。まずステップ2301としてデータ管理部101が文書DB107にユーザの指定した文書群を格納する。

【0085】続くステップ2302では、単語検出部102が文書群から単語キーワードを検出し、単語辞書105に格納する。こうして分類処理を行うためのデータが用意できると、ステップ2303で文書分類部103が図3の流れ図に従って分類体系を生成し、文書を分類する。分類結果出力部104はこの分類結果を図21のようなインタフェース画面に表示して、ユーザに提示する(ステップ2304)。ここまでは第1の実施例と同一の処理ステップである。さらに、分類結果を参照したユーザから分類指定部2201に対する指示を確認し

(ステップ2305)、指示がない場合には終了する。指示がある場合にはステップ2306で指示内容を解釈し、フォルダ数指定インタフェース2500への指示の場合はステップ2307の再分類処理を施し、分類体系構築補助インタフェース2700への指示の場合はステップ2308の再分類処理を適用する。フォルダ数指定インタフェース2500および分類体系構築補助インタフェース2700については後述する。そして再びステップ2304に戻り、再分類の結果をユーザに提示する。ユーザが分類指定部2201に対して指示するケースとしては、分類結果がユーザの意向に合わない場合などが挙げられる。

【0086】次に、分類指定部2201の提供する二つの入力画面、フォルダ数指定インタフェース2500および分類体系構築補助インタフェース2700について説明する。ここでは、各インタフェースから取り込まれる再分類に関する指示情報と再分類処理2307および2308の詳細について述べる。

【0087】はじめにフォルダ数指定インタフェース2500について述べる。

【0088】図24は分類結果出力部104によって作成された、約一千件の「コンピュータ」関係の文書群の分類結果を示すインタフェース画面である。これは図21と同種の出力画面であり、第1の実施例と同様に文書分類部103で作成された分類結果を読み出して作成される。具体的には図19の分類階層管理テーブル1900からフォルダ名1902、文書群1903、文書数1904、フォルダの階層関係(1905、1906、1907)を読み出すことによって、図24の画面に表示

2.0

する情報を得ている。

【0089】このように一つの分野の文書群であっても内容が多岐に渡れば詳細に分類することができ、数十から数百ものフォルダが生成される。分類結果として第一階層に数十個のフォルダが生成され、各フォルダの下位に同等数のフォルダが生成された場合を例に、本図では分類体系の上位三階層を表示している。フォルダ「パソコン、発売、販売、ソフト」2405およびフォルダ「発売、価格、販売、見込」2406はユーザが参照しようとして選択した状態を反転して示している。2401は第一階層に生成された複数個のフォルダの名前を縦方向に列挙し、2402はユーザから選択された第一階層2401のフォルダ「パソコン、発売、販売、ソフト」2405の下位(第二)階層のフォルダ名を、2403は選択された第二階層2402のフォルダ「発売、価格、販売、見込」2406の下位(第三)階層のフォルダ名を縦方向に表示している。本図では第一階層(top class)2401のフォルダ名が6個しか見えないが、実際にはスクロールバー2404を用いて画面をスクロールさせることで数十個のフォルダ名を参照できる。

【0090】なお第一階層内をさらに細分化し第二階層以下を作成する細分類処理(図17)について、第1の実施例では「上位フォルダの作成に利用していないキーワードを用いて分類を行う」場合を仮定したが、本例では上位フォルダの作成に利用したキーワードも用いて分類する場合について取り挙げている。そのため、上位階層のフォルダの作成に利用されたキーワードが下位階層のフォルダにも出現することがあり、例えば第一階層(top class)2401のフォルダ「パソコン、発売、販売、ソフト」2405に含まれているキーワード「パソコン」、「発売」、「販売」、「ソフト」は第二階層2402のフォルダ「発売、価格、販売、見込」2406とフォルダ「パソコン、開発、シリーズ、新製品」2407、フォルダ「ソフト、東京、複雑、成功」2408にも含まれている。

【0091】この分類結果では一階層のフォルダ数が多く、所望のフォルダを見つけにくい。一方、一階層のフォルダ数を少なくし階層を深くして細分類すると、所望の文書を見つけるまで手間がかかる。分類結果として適切なフォルダの数やフォルダの大きさは、分類対象の文書数や文書の内容の均質さによって異なってくる。さらに、分類結果の適否はこれを参照するユーザの視点によって異なることから、予め適切なフォルダ数や平均文書数を設定することは難しい。そこで、フォルダ数指定インタフェース2500は文書分類部103によって生成される分類結果のフォルダ数やフォルダの大きさを、ユーザによって指定できる環境を提供するという役割を担う。

【0092】図25に示すフォルダ数指定インタフェー

2 1

ス2500では、フォルダ数やフォルダの平均文書数といった分類結果に関する情報をユーザに提示し、好みに応じて適切なフォルダ数や平均文書数にまとめ直すための指示を受け取れるようにしている。一般的にフォルダ数を少なくすれば平均文書数は多くなるというように両者は連動する関係にあるが、ユーザが分類結果を評価する基準としてどちらを用いてもよいようにする。2501は一階層に生成されたフォルダ数を、2502は平均文書数を示す。フォルダ数操作バー2503はフォルダ数の増減を、平均文書数操作バー2504は平均文書数の増減をユーザが指定するためのグラフィカルユーザインタフェース (GUI) である。2505は文書数に応じた大きさに表現した円状の図形をフォルダと見立て、分類体系全体の状況を表示している。分類指定部2201はフォルダ数指定インタフェースを作成するために、分類結果出力部104から第一階層2401に関する情報を読み出し、フォルダ数を調べて2501に書き込み、各フォルダの文書数を調べてその平均値を2502に書き込み、各フォルダは文書数に比例した値を半径とする円状の図形として画面2505を作成する。

【0093】図25はユーザがフォルダ数操作バー2503を操作してフォルダ数の減少を指示した様子を示している。ユーザからの指示に従って再分類処理2307が実行されると、新たな分類結果が生成されフォルダ数指定インタフェースは2500から2510のように変わる。再分類処理2307の詳細は後述する。

【0094】再分類前にはフォルダ数2501は96個であったが、ユーザの指示を反映して、フォルダ数2511は30個に減少している。このように、ユーザが自分の参照しやすいレベルに分類結果を調節することが可能となる。

【0095】次に、分類指定部2201がフォルダ数指定インタフェース2500から取り込まれたユーザの指示に基づいて行う、再分類処理2307の詳細を図26を用いて説明する。

【0096】まず、分類指定部2201はステップ2601としてユーザからの指示内容を解釈する。「フォルダ数の減少」あるいは「平均文書数の増加」を指示するものでなければ、ステップ2602として分類処理2303における分類結果を見直し、よりフォルダ数が多く、平均文書数が少ない分類結果を再選択することを文書分類部103に指示する。

【0097】ここで文書分類部103が行う、分類結果の再選択処理2602について説明を加える。文書分類部103の行う分類処理2303は図3の流れ図に基づき、関連キーワード処理302を繰り返し適用して分類結果となるフォルダを生成し、これらのフォルダに文書を分類する。関連キーワード処理302とは、図9で示した流れ図に従い、関連のありそうなフォルダの組を統合することによって分類結果となるフォルダを作り出

2 2

す。フォルダに関する情報を記録しておくフォルダテーブル700 (図7) は、関連キーワード分類処理302が繰り返されるたびに更新され、関連キーワード処理302が終了 (ステップ303) した時点の情報が分類結果のフォルダとして採用される。

【0098】すなわち、関連キーワード分類処理302が繰り返されるたびにフォルダが統合されて、分類結果全体としてみるとフォルダ数が減少し、平均文書数が増加する。そこで、関連キーワード分類処理302のたびにフォルダテーブル700の情報を中間結果的なフォルダとして記録しておけば、後から分類結果よりも多いフォルダ数で、平均文書数の少ないフォルダを再選択することが可能となる。例えば第1の実施例の図16のフォルダ「動物とねずみとペットとABC」1607が分類結果のフォルダの一つと仮定すると、中間結果にはフォルダ「動物とペット」1605、フォルダ「ねずみとABC国」1606が該当する。中間結果は関連キーワード処理302を繰り返した回数分だけ存在するので、ユーザの指示した操作バーの増減レベルに応じて「より少ないフォルダ数」で「より多い平均文書数」の中間結果を選択することによって、ユーザからの指示に対応する。

【0099】一方ステップ2601で、ユーザから「フォルダ数の減少」および「平均文書数の増加」が指示された場合は、分類指定部2201が文書分類部103に対して「より少ないフォルダ数で、より平均文書数の多い分類結果を作り直すこと」を指示する (ステップ2603)。これに対して文書分類部103では、分類結果のフォルダをさらに統合することによって「より少ないフォルダ数で、より平均文書数の多い分類結果」を作る。分類処理2303で分類結果となるフォルダが生成されるのは、図3の流れ図のステップ303で関連キーワード分類処理が終了したと判断される場合、すなわち図9のステップ905でフォルダの統合がこれ以上不可能であると判定された場合である。そこでステップ2603では、ステップ905の統合判定条件となるしきい値 T_d および T_σ を調整して、フォルダの統合がさらに可能となるような設定を行う。第1の実施例の図17の説明で述べたように、しきい値 T_d は平均ベクトルとの距離が T_d 以上に離れている場合に統合不適と判断する指標なので、 T_d をより大きい値に設定し直す。しきい値 T_σ は平均ベクトルとの距離の分散値が T_σ 以上に大きい場合に統合不適と判断する指標なので、 T_σ をより大きい値に設定し直す。具体的な値はユーザの指示した操作バーの増減レベルに応じて、文書分類部103が決定する。これら準備の後、関連キーワード処理302を適用してさらなるフォルダの統合を行い、可能な限りフォルダの統合を繰り返す (ステップ303)。これによって先にユーザに提示した分類結果よりもフォルダ数が少なく、平均文書数が多い分類結果を生成できるので、この結果から分類結果の再選択処理2602を行う。

23

【0100】次に、分類指定部2201の提供するもう一つの入力画面である、分類体系構築補助インタフェース2700の詳細を述べる。

【0101】文書分類部103はシソーラスなどの情報を用いずに、自動的にフォルダを階層化するため、一般的な上位語下位語の概念と矛盾する階層関係を生成する場合がある。例えば、図20のフォルダ「犬」1802とフォルダ「柴犬」1806は正しい上下関係にあるが、逆転した場合には概念的に矛盾することになる。

【0102】この解決策として、シソーラス辞書106を用いて上位語下位語として不適当な関係となるフォルダの生成を禁止する方法が考えられる。しかし、本発明の文書分類装置が生成するフォルダ名はキーワードを統合した形式、フォルダ「猫と魚」1801のようになることが多い。そこで前例の「犬」と「柴犬」のように、1対1のキーワードの上下関係を調べるのでは対応し切れず、複数個対複数個のキーワードの上下関係を考慮しなければならない。このとき、例えばキーワードA、B、C、Dから上位フォルダ「A、B」、下位フォルダ「C、D」が作られたとすると、「キーワードAとCは上位語下位語として適当であるが、キーワードBとDは逆転関係にある」という場合には適否判断が行えないという問題が残

り、これでは不十分である。

【0103】分類体系構築補助インタフェース2700は、適切な階層構造を作るための補助情報をユーザから取り込み、これを用いて文書分類部103が分類処理を行えるようにするものである。例えば特許の明細書ならば「発明の名称」、「特許請求の範囲」といった特定の項目と各項目に書くべき内容が定められているが、これらが補助情報に相当する。各項目に書かれる文章の内容は明細書ごとに異なるが、専門性や一般性の度合には共通点があり、項目ごとにその度合が決まっているものと考えられる。例えば明細書の「発明の名称」や「発明の属する技術分野」といった項目には発明の前提条件や背景が書かれるので、他の項目に比べて一般性が高い。また、「課題を解決するための手段」や「発明の実施の形態」などの項目には発明の内容が記載されることから、専門性が高くなる。上位語となるキーワードは一般性が高く、下位語となるキーワードは下位に位置するほど専門性が高い。

【0104】したがって、一般性の高い内容の項目に出現するキーワードを上位階層のフォルダに、専門性の高い内容の項目に出現するキーワードを下位階層のフォルダの作成に利用することで適切な階層構造を構築しやすくなる。具体的には、一般性の高い「発明の名称」や「発明の属する技術分野」といった項目に出現するキーワードを上位階層の作成に、課題を解決するための手段」や「発明の実施の形態」といった項目に出現するキーワードを下位階層の作成に利用するというように階層ごとに分類に利用する項目を限定する。分類体系構築補助イ

24

ンタフェース2700は、こうした文書に含まれる項目とその項目に書かれる文章の専門性をユーザから容易に取り込むことができる。

【0105】図27は文書分類装置2200が特許の明細書を対象として分類処理を行ったときの分類体系構築補助インタフェース2700の表示例である。次に、分類指定部2201がこの分類体系構築補助インタフェース2700を介してユーザから補助情報を取り込む処理の流れを図28の流れ図を用いて説明する。分類指定部2201は、ステップ2801として画面2701にサンプル文書を読み込む。サンプル文書とは文書DB107に格納されている文書群のうちのひとつで、図27では特許の明細書の一例である。続くステップ2802ではユーザからサンプル文書中の項目に関する情報を受け取る。ユーザはマウス2702を用いて画面2701上の文字列を指定することができる。図27ではマウス2702を用いて項目「発明の名称」の文字列をドラッグし、項目として指定した様子を示している。このようにユーザから項目が指定されるとステップ2803としてダイアログボックス2706を表示し、ユーザの指定した項目が適当かどうかの確認を求める。ステップ2804でユーザから確認が取れると、項目とその項目のサンプル文書上の出現位置に関する情報を取り込む(ステップ2805)。すなわち、分類指定部2201は「発明の名称」という項目と出現位置「1文字目から5文字目まで」という情報を格納する。サンプル文書中の全項目の受取りが終了すると、ステップ2806として各項目の出現位置を手掛かりとして項目をサンプル文書での出現順に整列し、これを項目設定画面2710の2712に表示する(ステップ2707)。続くステップ2808では、ユーザから全項目の専門性の度合に関する情報を受け取る。分類指定部2201は予め用意しておいた専門性を示す数種類の度合をレベルリスト2714に表示するので、これを用いてユーザは2712から項目の一つを選択し、その内容の一般性、専門性を考慮して2714から適切なレベルを一つ選択し、2715のOKボタンで確定するという手順で指定が行える。項目設定画面2710では「一般的」、「やや一般的」、「どちらともいえない」、「やや専門的」、「専門的」といった項目に関する5種類の専門性の度合が用意されているので、各項目に適切な度合をここから選択する。項目に関する専門性の度合のことを項目レベルと呼び、詳細については後述する。

【0106】次に図29を用いて、分類指定部2201が分類体系構築補助インタフェース2700から取り込まれたユーザの指示に基づいて行う、再分類処理2308の詳細を説明する。ステップ2901として分類指定部2201はユーザから取り込んだ項目とその項目レベルに基づいて階層構築情報を作成する。階層構築情報とは、文書分類部103が適切な階層構造の分類結果を構

25

築するために分類処理2303で参照する情報であり、ある階層を構築する場合に分類に利用すべき項目とその項目の重要度を規定するものである。詳細については後述する。

【0107】次にステップ2902として、この階層構築情報を用いて一時キーワードDBを作成する。一時キーワードDBとは、ある階層を構築する場合に分類に利用するキーワードとしての重要度を付与したキーワード群の集合を格納したもので、第1の実施例のキーワードDB108と同形式である。一時キーワードの作成に関しては後に説明する。これらは分類指定部2201がユーザから取り込んだ情報をもとに、適切な階層構造の分類結果を構築するためのデータを準備する処理ステップである。

【0108】次に、分類指定部2201は文書分類部103に対してこれらのデータを用いて適切な階層構造の分類結果を構築するように指示する。文書分類部103はステップ2903として、第1の実施例におけるキーワードDBの代わりに一時キーワードDBを用いて、各キーワードをキーワード群中に含む文書を各キーワードごとにまとめる、単一キーワード処理301を行う。

【0109】次にステップ2904として、単一キーワード分類処理2903によって作成された単一キーワードフォルダについて、類似した文書群を含む単一キーワードフォルダの組を統合する関連キーワード処理302を行う。第2の実施例では、第1の実施例で説明した図14のステップ1402における統合すべきかどうかを判定するための処理において、適切な階層構造の分類結果を構築するために階層構築情報を用いてワードベクトルの各キーワードに重要度を付与する。この詳細は後述する。

【0110】続くステップ303では第1の実施例と同様に、関連キーワード処理2904によって作成された関連キーワードフォルダについて、類似した文書群を含む関連キーワードフォルダの統合を繰り返す。さらに、作成された単一キーワードフォルダや関連キーワードフォルダ内について、細分類できるかどうかを調べ（第1の実施例のステップ304と同様）、細分類可能な場合はフォルダ内を分類する（ステップ2905）。ステップ2905はフォルダ内に分類されている文書を対象として、図29におけるステップ2902から終了までの処理Bを適用する。ステップ306では、すべてのフォルダ内について階層的な分類を繰り返す行。

【0111】次に、階層構築情報（の詳細とその役割について説明する。ここでは、文書分類装置2200は3階層の分類体系を生成するものとし、項目レベルは「一般的」、「やや一般的」、「どちらでもない」、「やや専門的」、「専門的」の5種類を設定できるようにするものとする。さらに図28の流れ図に従い、分類体系構築補助インタフェース2700を介して、図30の4項

26

目と各項目レベルがユーザから与えられている。階層構築情報を作成するための規則として、例えば次のようなものを仮定する。

【0112】○第一階層構築規則：「一般的」項目レベルの重要度 = a

「やや一般的」項目レベルの重要度 = b

その他の項目レベルの重要度 = 0

○第二階層構築規則：「やや一般的」項目レベルの重要度 = c

10 「どちらでもない」項目レベルの重要度 = d

「やや専門的」項目レベルの重要度 = e

その他の項目レベルの重要度 = 0

○第三階層構築規則：「やや専門的」項目レベルの重要度 = f

「専門的」項目レベルの重要度 = g

その他の項目レベルの重要度 = 0

例えばこの第一階層構築規則は、第一階層のフォルダを生成する際には「一般的」項目レベルの重要度をa、

20 「やや一般的」項目レベルの重要度をbとし、その他の項目レベルは重要度0、すなわち分類に利用しないことを意味する。例えば、第一階層は一般的な内容に基づき分類するのが好ましいと考え、と、「一般的」項目レベルの重要度aを1、「やや一般的」項目レベルの重要度bを0.5、その他の項目レベルの重要度を0とするというようにa~gには0以上の1以下の定数を経験的に決めて、分類指定部2201に設定してあるものとする。

【0113】上記規則に基づき、図30の項目レベルから階層構築情報を作成すると以下ようになる。

【0114】○第一階層構築情報：(a, 0, 0, b)

30 ○第二階層構築情報：(0, e, 0, c)

○第三階層構築情報：(0, f, g, 0)

階層構築情報の第一要素は項目「第一章」、第二要素は項目「第二章」、第三要素は項目「第三章」、第四要素は項目「第四章」それぞれの重要度である。すなわち、上例の第一階層構築情報は項目「第一章」を重要度a、項目「第四章」を重要度bとして分類に利用し、それ以外の項目は分類には利用しないことを示す。

【0115】次に、一時キーワードDBの詳細について説明する。第一階層を構築する場合のステップ2902では、第一階層構築情報を用いて次のように一時キーワードDBを作成する。

【0116】まず、第一階層構築情報から分類に利用すべき項目は「第一章」と「第四章」であることを読み取る。すなわち、「第二章」と「第三章」は重要度0なので分類に利用せず、それ以外の「第一章」と「第四章」を利用する。

【0117】次に、キーワードDB108から文書のキーワード群を読み出し、項目「第一章」と項目「第四章」に出現するキーワードだけを取り出して一時キーワード群を作成し、これを一時キーワードDBに登録する。例え

27

ば、図31の文書3100のキーワード群はキーワードDB108中のキーワード群3110として登録されている。キーワード群3110から一時キーワード群を作るには、項目「第一章」3101と項目「第四章」3103にそれぞれ付随する文章3102、3104中に出現するキーワードだけを取り出せばよい。すなわち、キーワードa3105、キーワードb3106、キーワードg3107、キーワードh3108がこれに相当し、一時キーワード群3112を作成し、一時キーワードDB3111に登録される。単一キーワード分類2903では

10 こうして作成された一時キーワード群を利用する。
 【0118】こうした階層構築情報、一時キーワードDBを用いて行う関連キーワード分類処理2904について説明を加える。第一階層を構築するための関連キーワード処理2904では、第一階層構築情報から項目「第一章」に出現するキーワードは重要度a、項目「第四章」に出現するキーワードは重要度bであることを読み取る。次に、各文書の持つキーワードの出現頻度に上記重要度を積算し、第一階層構築時に重視すべきキーワードの重要度を高めて分類処理を行う。このように、ユーザの指定した項目に出現するキーワードを重視して分類

20 することでその階層に適するフォルダが生成しやすくなり、上下関係の適切な分類体系が作り出されることになる。
 【0119】第2の実施例で説明した分類指定部2201について、予めユーザから分類に関する指示を受け取りこれに基づく分類処理を施す、文書分類装置4100（図41に示した）に関する第3の実施例を説明する。

【0120】ユーザがキーボード110、マウス111といった入力装置を用いて分類したい文書群を指定し、分類処理の実行を指示すると、文書分類装置4100が起動され、図32の流れ図に基づく処理が施される。まずステップ2301としてデータ管理部101が文書DB107にユーザの指定した文書群を格納する。

【0121】続くステップ2302では、単語検出部102が文書群から単語キーワードを検出し、単語辞書105に格納する。こうして分類処理を行うためのデータが用意できると、ステップ2305として分類指定部2201に対するユーザからの指示を確認し、指示がない場合には待ち続ける。ユーザから指示があると文書分類部103にユーザの指示を受け渡して分類処理3201を実行させる。

【0122】分類処理3201の詳細は後述する。次にステップ2304として分類結果出力部104が分類結果を表示する。さらに分類指定部2201は分類結果を参照したユーザからの指示を確認し（ステップ2305）、指示がない場合には処理を終了する。指示がある場合にはステップ2306で指示内容を解釈し、フォルダ数指定インタフェース2500への指示の場合はステップ2307の再分類処理を施し、分類体系構築補助イ

28

ンタフェース2700への指示の場合はステップ2308の再分類処理を適用する。ステップ2305で再びユーザからの分類に関する指示がある例としては、予めユーザが指定した分類結果が予想に反して意向に沿わなかった場合などが考えられる。

【0123】図33に分類処理3201の具体的な処理の流れを示す。まずステップ3301として、フォルダ数指定インタフェース2500に対する指示か、分類体系構築補助インタフェース2700に対する指示かを調べ、後者の場合は分類体系構築補助インタフェース2700の再分類処理2308を実行して処理を終了する。前者の場合はステップ2303で文書分類部103が図3の流れ図に従って分類体系を生成し文書を分類する。次にステップ3302として、分類結果がユーザの指示条件を満たすかどうかを確認する。フォルダ数指定インタフェース2500へのユーザからの指示はフォルダ数あるいは平均文書数が指定されるものなので、これに見合った分類結果が得られているかどうかを調べ、得られた場合は処理を終了する。ここでユーザからの指示が満たされない場合は、フォルダ数指定インタフェース2500の再分類処理2307を実行してユーザの指示に見合うように再び分類処理2303を実行する。

【0124】図1に示した第1の実施例の文書分類装置100における分類結果出力部104に、ユーザが分類結果をブラウジングする手段を設けた文書分類装置に関する第4の実施例を説明する。

【0125】第1の実施例の文書分類装置は予め設定した分類体系に基づいて分類する場合と異なり、分類結果として生成されるフォルダが全く未知である。そこでユーザが分類結果を参照して所望の文書を検索する場合には、まずどのようなフォルダがあるのかを調べ、その上で所望の文書が分類されているようなフォルダを選びフォルダ内をブラウジングする。しかし、所望の文書に辿り着くまでの手間が大きいと分類体系が固定化されている方がブラウジングしやすいということになる。これでは分類対象の文書に応じて適切な分類体系を生成する本発明が有効に活用されない。そこで図34のように、第1の実施例の文書分類装置に検索支援部3401を設けた文書分類装置3400の構成を取り、分類結果のブラウジングの支援が行えるようにする。検索支援部3401はフォルダ検索支援機能3402と文書検索支援機能3403という二つの処理機能から構成される。

【0126】ユーザがキーボード110、マウス111といった入力装置を用いて分類したい文書群を指定し、分類処理の実行を指示すると、文書分類装置3400が起動され、図35の流れ図に基づく処理が施される。まずステップ2301としてデータ管理部101が文書DB107にユーザの指定した文書群を格納する。

【0127】続くステップ2302では、単語検出部102が文書群から単語キーワードを検出し、単語辞書1

29

05に格納する。こうして分類処理を行うためのデータが用意できると、ステップ2303で文書分類部103が図3の流れ図に従って分類体系を生成し、文書を分類する。分類結果出力部104はこの分類結果を図36のようなインタフェース画面3600に表示して、ユーザーに提示する(ステップ2304)。

【0128】本図では3601に第一階層のフォルダ名が、3602に第二階層のフォルダ名が、3603に第三階層のフォルダ名が表示されている。ここまでは第1および第2の実施例と同一の処理ステップである。続くステップ3501では、分類結果出力部104が表示した分類状況のうち、任意のフォルダあるいは文書をユーザーが選択したかを調べる。もし何も選択がないまま、ユーザーが終了ボタン3604で分類結果の終了を指示した場合は処理を終える(ステップ3502)。

【0129】選択した場合は、ステップ3503で分類結果出力部104がフォルダを選択しているのか文書を選択しているのかを調べ、フォルダの場合は検索支援部3401のフォルダ検索支援機能3402にフォルダ検索支援処理3504を指示し、文書の場合は検索支援部3401の文書検索支援機能3403に文書検索支援処理3505を指示する。フォルダ検索支援処理3504および文書検索支援処理3405はブラウジングしやすいように分類結果の加工を行うもので、詳細については後述する。さらに、ユーザーが検索支援ボタン3605を押し、分類結果のブラウジングの支援を要求した場合には、検索支援部3401が分類結果出力部104に加工した分類結果を受け渡し、新たな分類結果が表示される。

【0130】次に、検索支援部3401の提供する、フォルダ検索支援機能3402および文書検索支援機能3403の詳細について述べる。

【0131】フォルダ検索支援機能3402とは、ユーザーが参照しようとして選択したフォルダとの類似度に基づいてその他のフォルダを評価し、この類似度順にフォルダの並び換えを行うものである。図36は分類結果出力部104が出力する分類結果のインタフェース画面3600と、検索支援部3401が出力する補助情報画面3610を示している。インタフェース画面3600には、第一階層3601、第二階層3602、第三階層3603のフォルダ名が表示され、ユーザーが選択し参照できるようにになっている。ここでは説明を簡略化するため、フォルダ名を抽象化して「フォルダa」などのように示しているが、実際には1個以上のキーワードによって構成される。補助情報画面3610には、フォルダ数3611や平均文書数3612といった分類結果に関する補助情報が表示される。ユーザーがまだ何も選択していない状態のときは、画面3613には第一階層に生成されたフォルダがその文書数に応じた大きさの円で表現され、文書数の多い順番に整列される。ユーザーはこれらの

30

情報を手掛かりとして所望のフォルダを選択し、フォルダに分類された文書を参照する。

【0132】例えば、ユーザーが第一階層3601の中からフォルダaを選択すると、分類結果出力部104はこれを検索支援部3401のフォルダ検索支援機能3402に知らせ、フォルダ検索支援処理3504を実行する。フォルダ検索支援処理3504では、まずフォルダaと第一階層のその他のフォルダとの類似度を調べる。フォルダ間の類似度は、「フォルダaに分類された文書と一致する文書が多いフォルダほど類似度が高い」と判断する。これは第1の実施例で説明した文書分類部103の関連キーワード処理302の中で求められる。すなわち、第4の実施例では図35のステップ3501の分類処理がこれに相当する。

【0133】フォルダ検索支援処理3504ではこのフォルダ間の類似度が必要になるので、第4の実施例の分類処理3501ではフォルダ間で一致する文書数に関する情報を保存しておき、フォルダ検索支援機能3402に情報を提供する。具体的には図9のステップ902でフォルダ間一致文書数管理テーブル1001の内容の保存を行う。

【0134】例えば、フォルダaとその他のフォルダとの類似度を調べるには、まずフォルダ間一致文書数管理テーブル1001を参照して、フォルダ識別番号1002、1003からフォルダaに対応する識別番号のカラムだけを取り出す。フォルダ間一致文書数管理テーブル1001は一致文書数を手掛かりとして降順に整列されているので、取り出したカラムのうち、テーブルの上位に位置するカラムほどフォルダaとの一致文書数が多く、類似度が高いことになる。すなわち、フォルダaとの類似度順に第一階層のフォルダに関する情報を取り出すことができる。フォルダ検索支援機能3402はこれらの情報を用いて、補助情報画面3610の画面3613を図37の補助情報画面3700の画面3701のように書き換える。すなわち、フォルダaと類似する順番に第一階層のフォルダを並び換え、ユーザーがフォルダaと類似したフォルダに関する情報を得やすくなるようにしている。

【0135】次に、文書検索支援機能3403について説明する。

【0136】文書検索支援機能3403とは、フォルダに分類された文書のうち、ユーザーが不適切と判断した文書と類似する文書を調べ、これをフォルダから除外することで所望の文書を検索する操作を支援するものである。図38は分類結果出力部104が出力する分類結果のインタフェース画面3800と、検索支援部3401が出力する補助情報画面3810を示している。インタフェース画面3800には、第一階層3801にフォルダ名、3802にはユーザーによって選択されたフォルダaに関する第二階層のフォルダ名、3803にはサブフ

31

フォルダaに関する文書名一覧が表示され、ユーザが3803から文書aを選択し、文書の内容を表示する画面3804で文書aを参照している。

【0137】補助情報画面3810の3811には、現在ユーザが参照中の文書aをはじめとするサブフォルダa中の文書と、各文書の内容の先頭部分が一覧表示されている。ユーザはこれらの情報を手掛かりとして分類結果をブラウジングする。情報を参照した結果、例えば「文書aは不要」とユーザが判断し、不要な文書を指定する消去ボタン3812を押すと、検索支援部3401はこの情報を文書検索支援機能3403に知らせる。文書検索支援機能3403は、図39の流れ図に従って文書検索支援処理3505を行う。不要と判断された文書aと類似する文書を調べるために、第1の実施例で述べたワードベクトル間の距離計算処理1405を用いる。そのための準備として、フォルダ内文書の距離計算処理1300をサブフォルダaについて行う。

【0138】次に、図14のステップ1401でワードベクトルテーブル1500の初期化を行い、各文書についてワードベクトルの作成を繰り返す（ステップ1402、1403）。ここまでは第1の実施例と同一の処理である。続くステップ3901では、文書aとその他の各文書とについて第1の実施例のワードベクトル間の距離計算処理1405を行う。文書間の距離は類似度であることから、文書aと各文書との類似度が求められる。さらにステップ3902として、文書aとの距離がしきい値 $T\alpha$ 以下の文書、すなわち文書aとの類似度が近いものを調べ、これらを不適切な文書の候補とみなす（ステップ3903）。全文書についてこの処理を繰り返す（ステップ3904）、文書aと類似度の高いすべての

文書を選び出す。

【0139】検索支援部3401は、不適切な文書の候補と判断された文書に関する情報を分類結果出力部104に受け渡す。この後、ユーザが検索支援ボタン3805を指定すると、文書aと類似した文書をフォルダから除外して、新たに図40の分類結果のインタフェース画面4000を作成する。サブフォルダaは文書aおよび文書aと類似すると判断された文書が除かれて文書数が図38の96件（3806）から図40の71件（4001）になり、文書一覧3803は文書aおよび文書aと類似すると判断された文書d、文書fなどの25件が除かれて4002のように変更されている。

【0140】

【発明の効果】本発明によれば、以下の効果が得られる。

【0141】（1）既定の分類体系に依存することなく、文書を分類することができる。

【0142】（2）階層的な分類体系を自動的に生成することができる。

【0143】（3）分類結果群の代表名称を付与するこ

32

とができる。代表名称は分類結果群との関連度の高いものから順番に付与する。

【0144】その結果、ユーザが特に興味の対象を指定しなくても、大量の文書の中から所望の文書を見つけやすくなることができる。あるいは興味の分野は定まっていなくても、それを代表するキーワードが思いつかない場合の助けとなる。

【図面の簡単な説明】

【図1】本発明の一実施例のシステム構成図である。

【図2】本実施例のキーワード説明図である。

【図3】本実施例の文書分類処理の流れ図である。

【図4】本実施例の単一キーワード分類処理の流れ図である。

【図5】本実施例のキーワードテーブルのデータ構造を表わす図である。

【図6】本実施例のキーワードテーブルの具体例を示す図である。

【図7】本実施例のフォルダテーブルのデータ構造を表わす図である。

【図8】本実施例の単一キーワード分類処理の具体例を示す図である。

【図9】本実施例の関連キーワード分類処理の流れ図である。

【図10】本実施例のフォルダ間一致文書数管理テーブルのデータ構造を表わす図である。

【図11】本実施例のフォルダ間一致文書数管理テーブルの具体例を示す図である。

【図12】本実施例のフォルダ統合時に生じるフォルダ間一致文書数管理テーブルの更新処理の具体例を示す図である。

【図13】本実施例のフォルダ内文書情報の距離計算の流れ図（その1）である。

【図14】本実施例のフォルダ内文書情報の距離計算の流れ図（その2）である。

【図15】本実施例のワードベクトルテーブルの具体例を示す図である。

【図16】本実施例の関連キーワード分類処理結果の具体例を示す図である。

【図17】本実施例の細分類処理の流れ図である。

【図18】本実施例の細分類処理結果の具体例を示す図である。

【図19】本実施例の分類階層管理テーブルのデータ構造を表わす図である。

【図20】本実施例の最終的な分類結果例を示す図である。

【図21】本実施例の分類結果の画面表示の具体例を示す図である。

【図22】本発明の第2の実施例のシステム構成図である。

【図23】第2の実施例の文書分類装置の流れ図であ

33

る。

【図24】第2の実施例の分類結果の画面表示の具体例を示す図である。

【図25】第2の実施例のフォルダ数指定インタフェースの具体例を示す図である。

【図26】第2の実施例のフォルダ数指定インタフェースにおける再分類処理の流れ図である。

【図27】第2の実施例の分類体系構築補助インタフェースの具体例を示す図である。

【図28】第2の実施例の分類体系構築補助インタフェースにおいて項目設定を行う画面の具体例を示す図である。

【図29】第2の実施例の分類体系構築補助インタフェースにおける再分類処理の流れ図である。

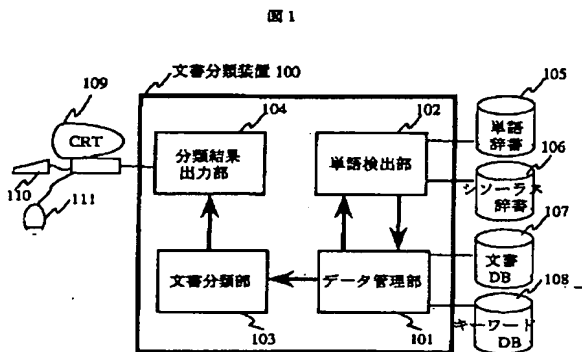
【図30】第2の実施例の項目と項目レベルの具体例を示す図である。

【図31】第2の実施例の一時キーワードDBの説明図である。

【図32】第2の実施例の分類指定部2201を図23とは別に適用した場合の、文書分類装置の流れ図である。

【図33】第2の実施例の図32における分類処理の詳

【図1】



【図10】

図10

フォルダ間一致文書数管理テーブル1001

フォルダ識別番号	フォルダ識別番号	一致文書数	統合済フラグ
0	1	6	済

34

細を示す図である。

【図34】第3の実施例のシステム構成図である。

【図35】第3の実施例の文書分類装置の流れ図である。

【図36】第3の実施例の分類結果の画面表示の具体例およびフォルダ検索支援機能によって提示される補助情報画面の具体例を示す図である。

【図37】第3の実施例の補助情報画面がユーザからの指示に基づいて変化した具体例を示す図である。

【図38】第3の実施例の分類結果の画面表示の具体例および文書検索支援機能によって提示される補助情報画面の具体例を示す図である。

【図39】第3の実施例の文書検索支援機能による検索支援処理の流れ図である。

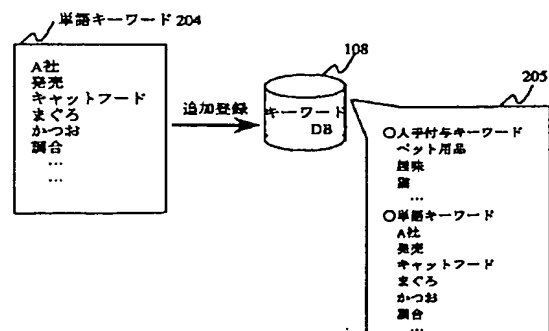
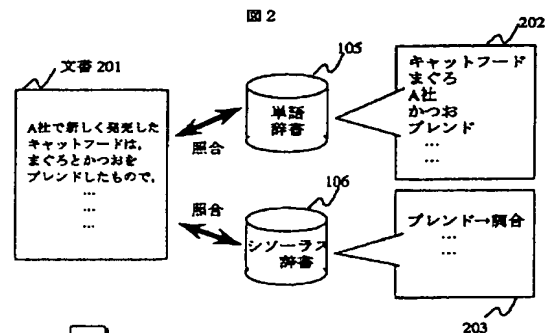
【図40】第3の実施例の文書検索支援機能によって分類結果が変化した具体例を示す図である。

【図41】第3の実施例のシステム構成図である。

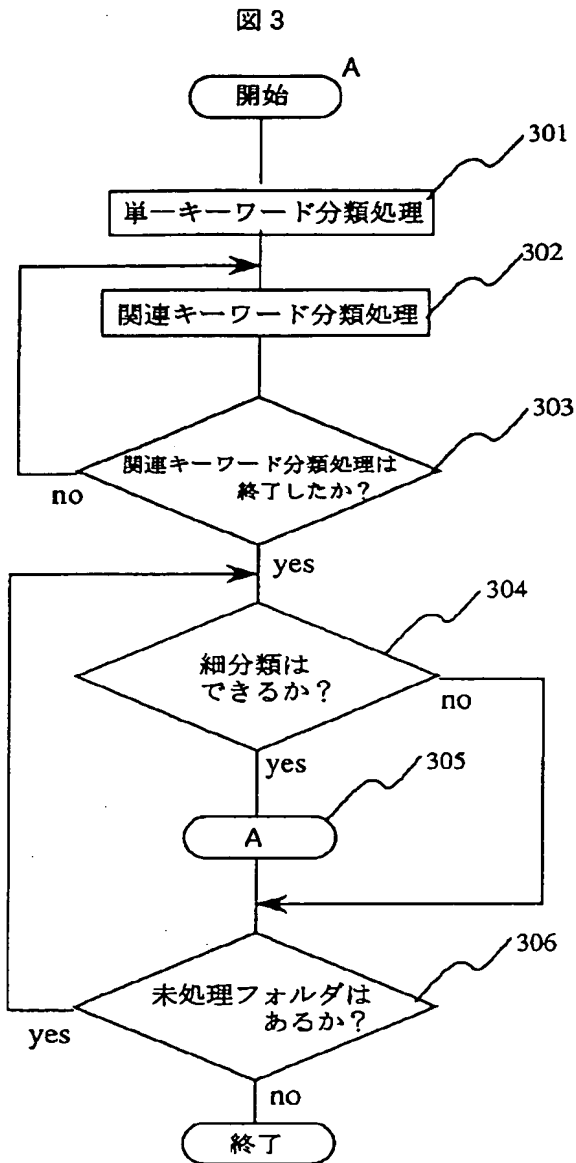
【符号の説明】

100…文書分類装置、102…単語検出部、103…文書分類部、301…単一キーワード分類処理、302…関連キーワード分類処理、1300…フォルダ内文書の距離計算

【図2】



【図3】



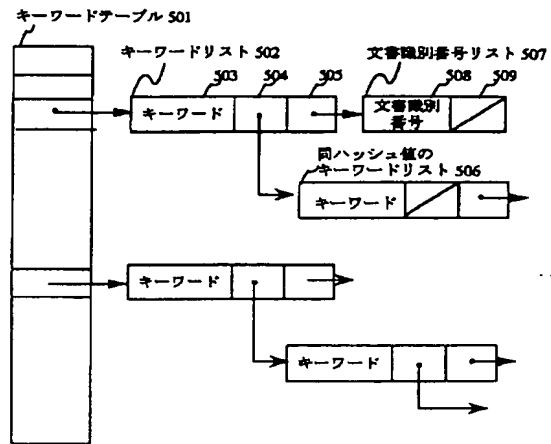
【図30】

図30

項目	項目レベル
第一章	一般的
第二章	やや専門的
第三章	専門的
第四章	やや一般的

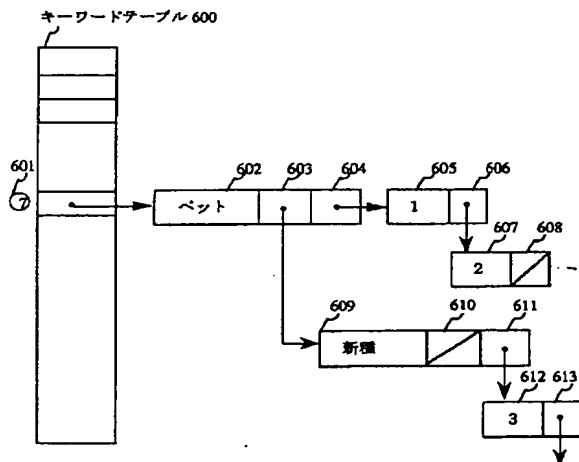
【図5】

図5



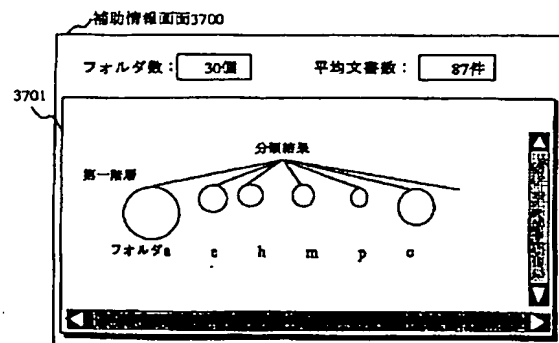
【図6】

図6



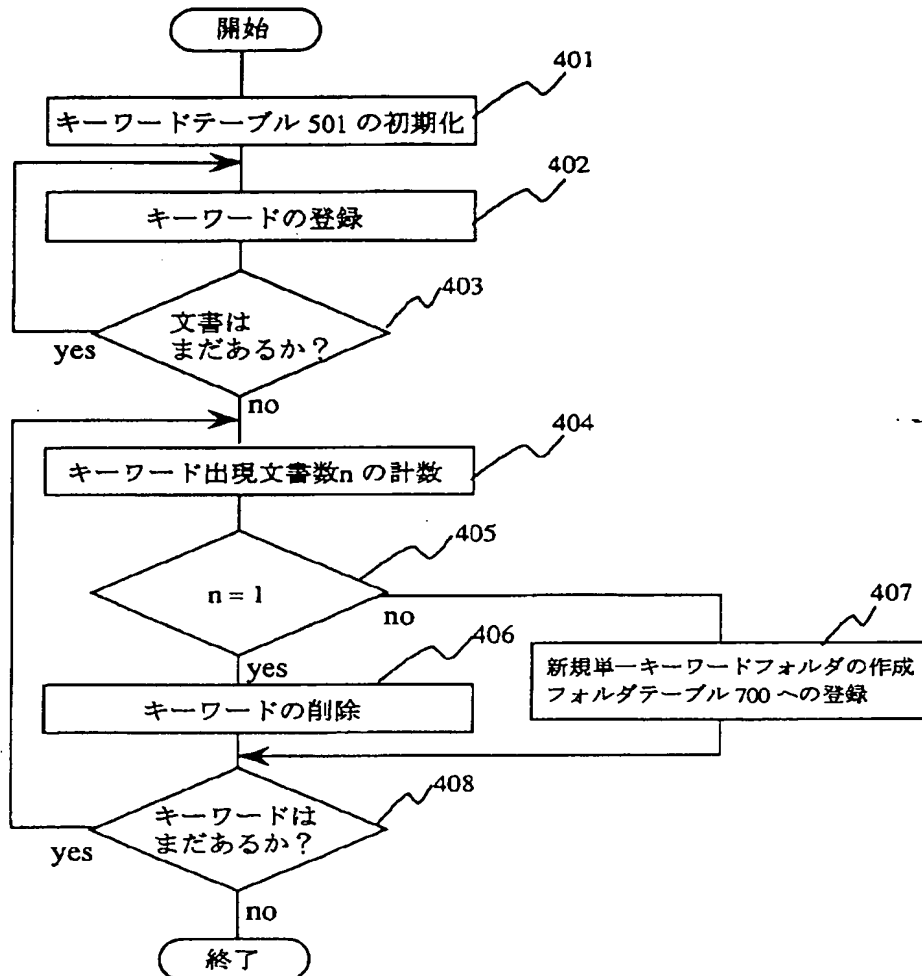
【図37】

図37



【図4】

図4



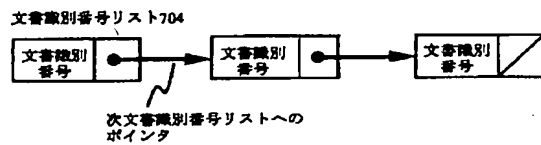
【図7】

図7

フォルダテーブル 700

701	702	703	704
フォルダ名	文書数	分類終了フラグ	文書識別番号リスト
ペット	2	OFF	→

705
①
フォルダ識別番号



【図11】

図11

文書数 1103

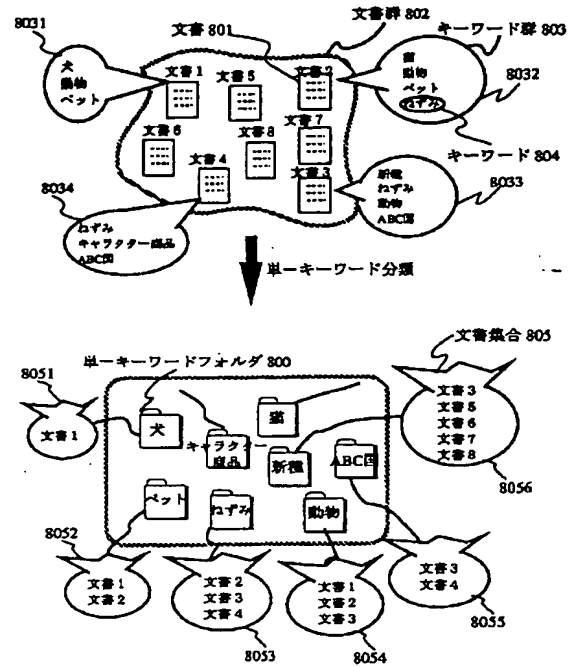
	0	1	2	3	4
0	0	10	9	9	8	
1	10		8	7	5	
2	9	8		2	1	
3	9	7	2		1	
4	8	5	1	1		
...						

← フォルダ識別番号 1102

↑ フォルダ識別番号 1101

【図8】

図8



【図12】

図12

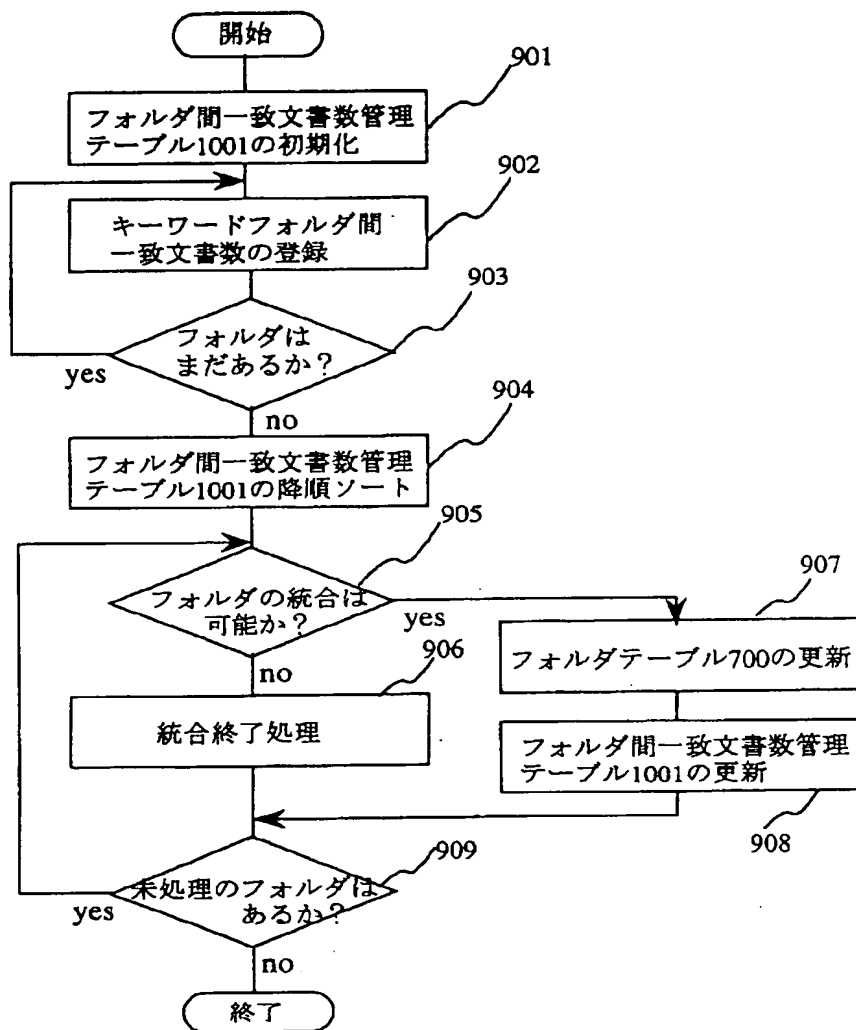
フォルダ間一致文書数管理テーブル 1200

1201	1202	1203	1204
フォルダ識別番号	フォルダ識別番号	一致文書数	統合済フラグ
0	0	1	10
1	3	5	9
2	5	1	9
3	0	2	8
4	0	3	8
5	2	5	8
6	5	3	7
7	3	1	5
8	0	4	2
9	1	2	2
10	6	5	1
11	2	4	1
12	1	4	1
13	0	5	1

1205 1214 1206 1207 1208 1209 1210 1211 1212 1213

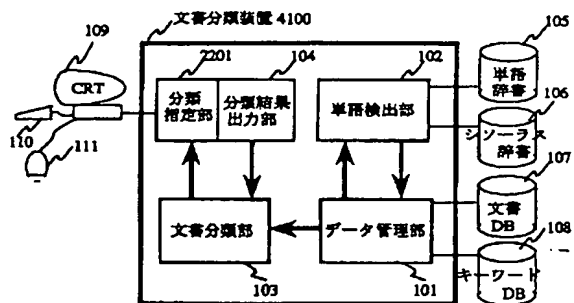
【図9】

図9



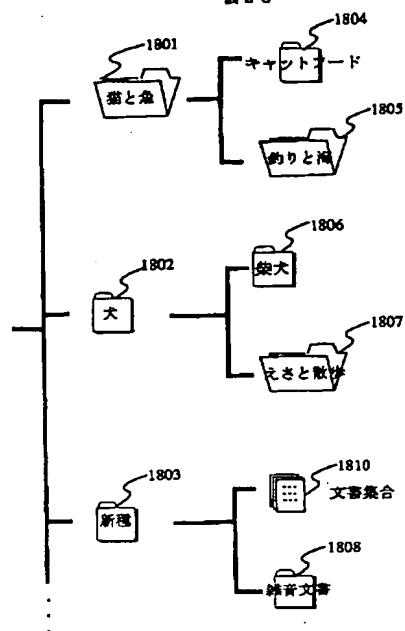
【図41】

図41



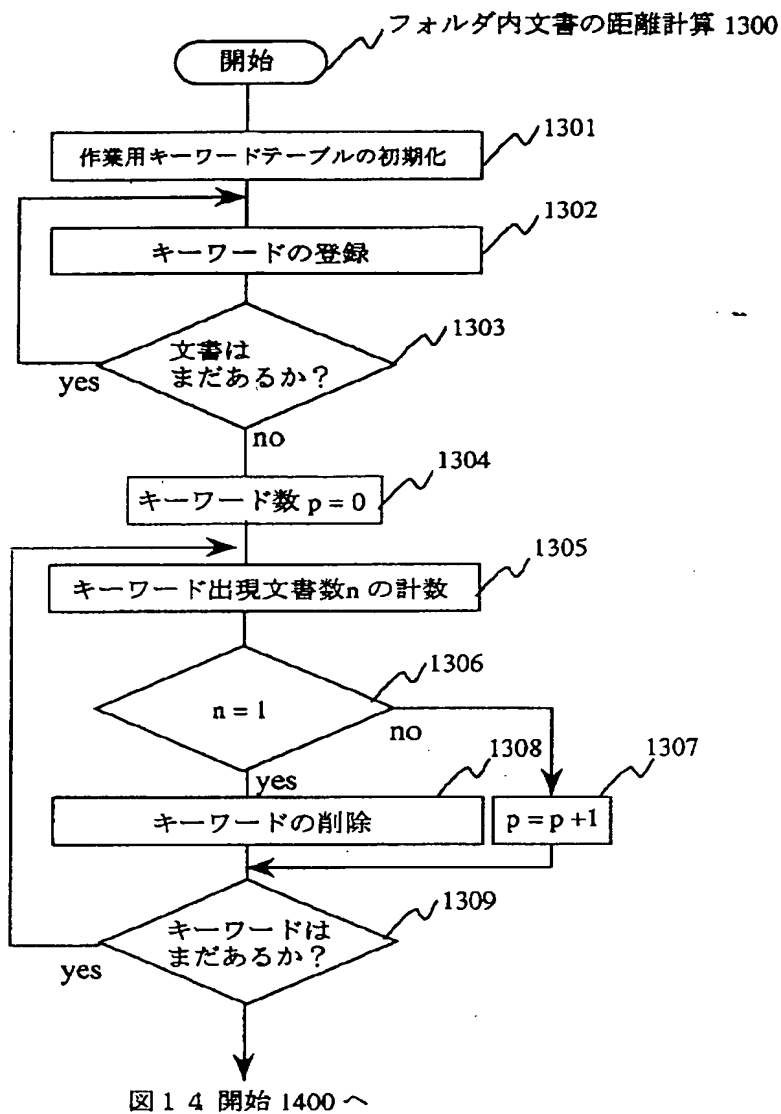
【図20】

図20



【図13】

図13



【図15】

図15

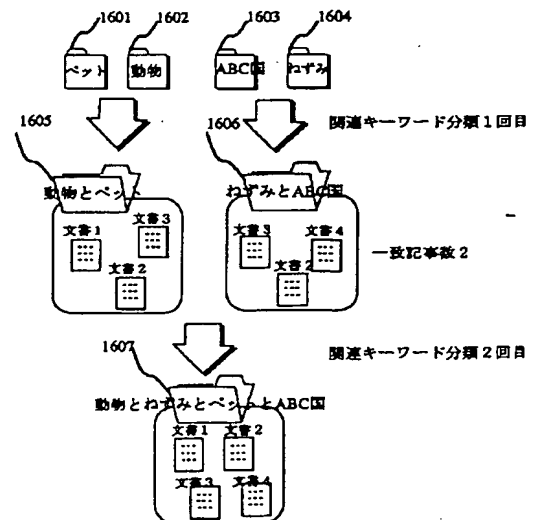
ワードベクトルテーブル 1500

	0	1	2	3	...	← キーワード識別番号 1502
0	3	2	1	1		
1	1	13	2	0		
2	1	1	8	0		
3	0	1	3	②		
...						
↑ 文書識別番号 1501						

文書中の出現頻度 1503

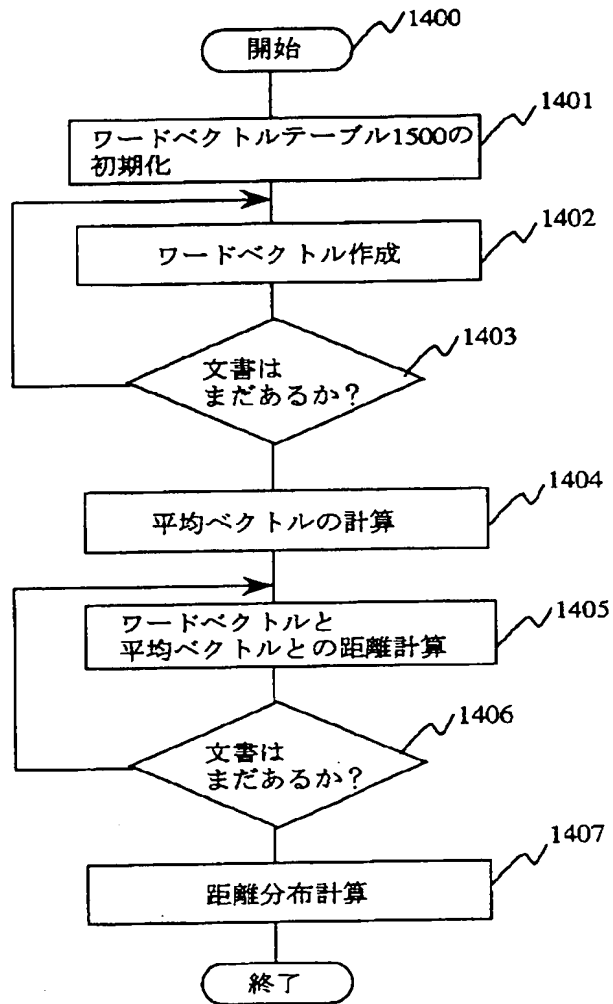
【図16】

図16



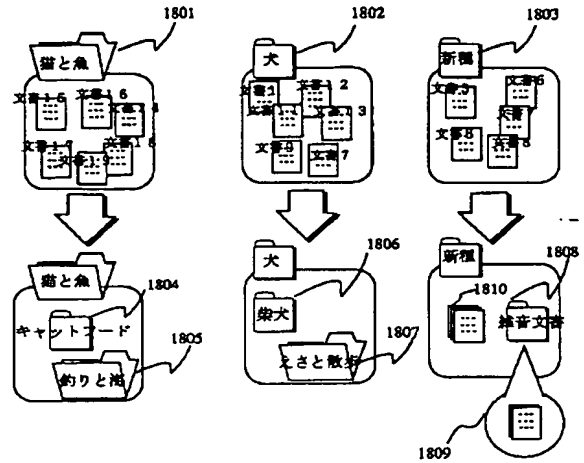
【図14】

図14



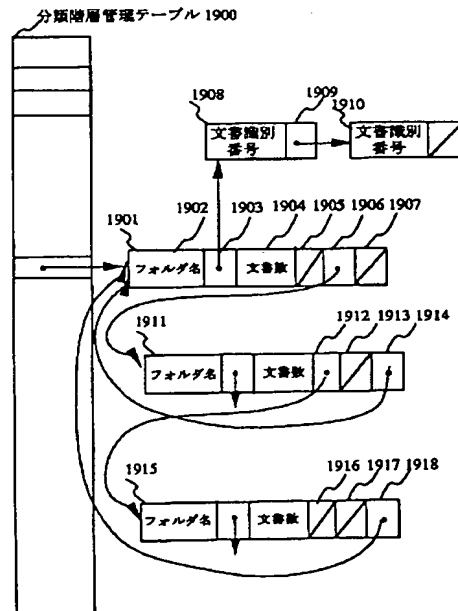
【図18】

図18



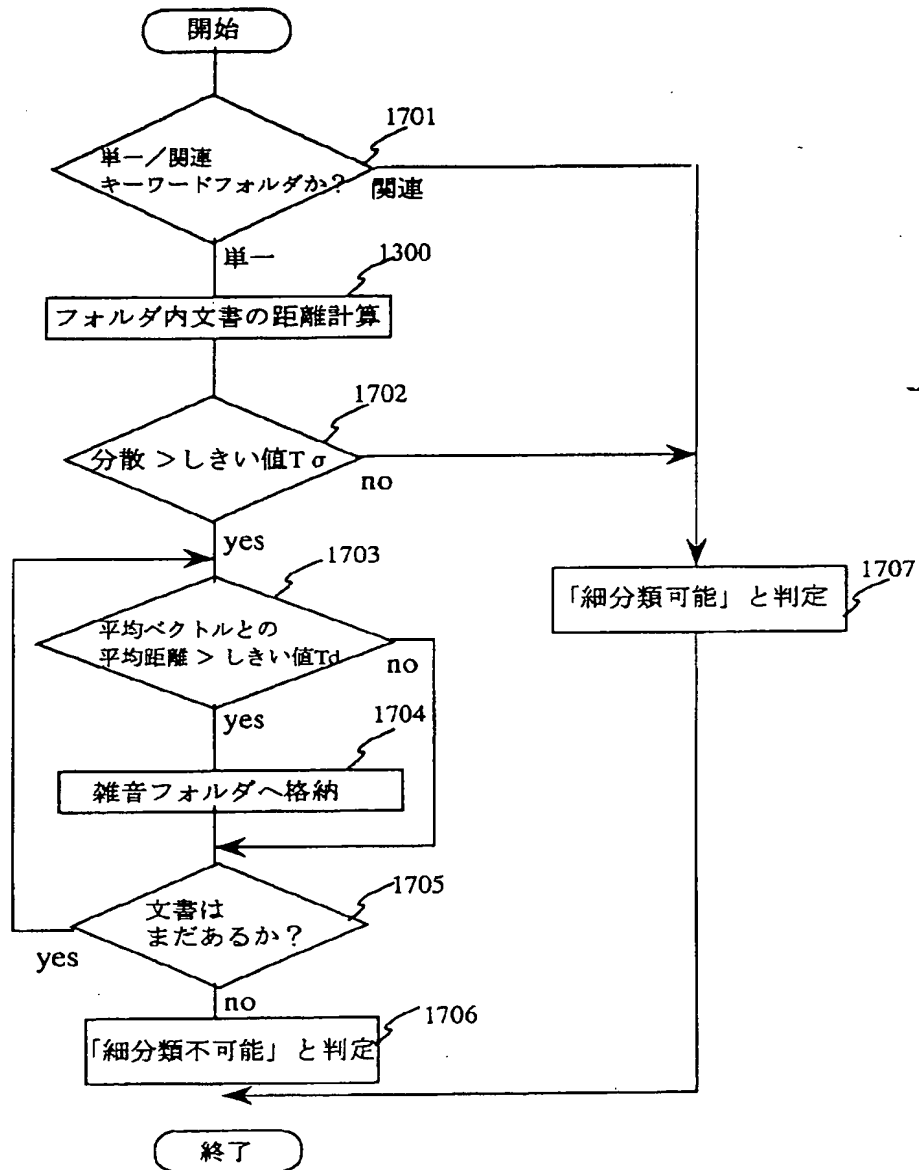
【図19】

図19



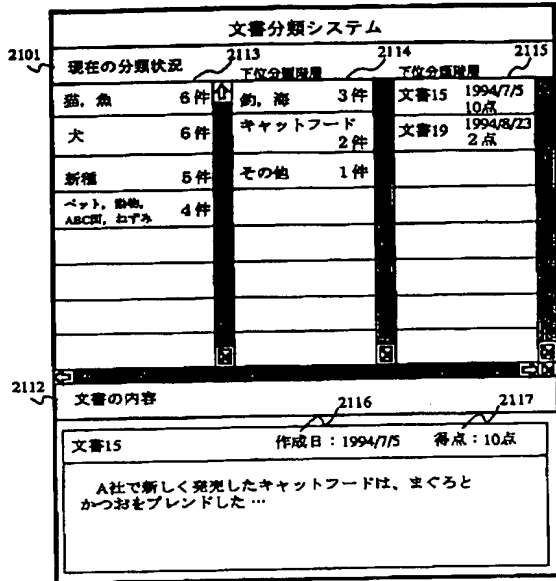
【図17】

図17



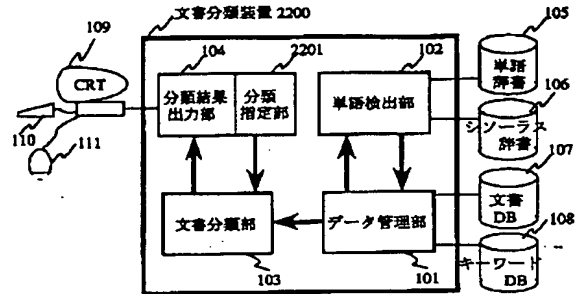
【図 2 1】

图 21

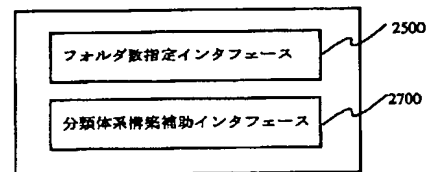


【圖 22】

圖 2 2



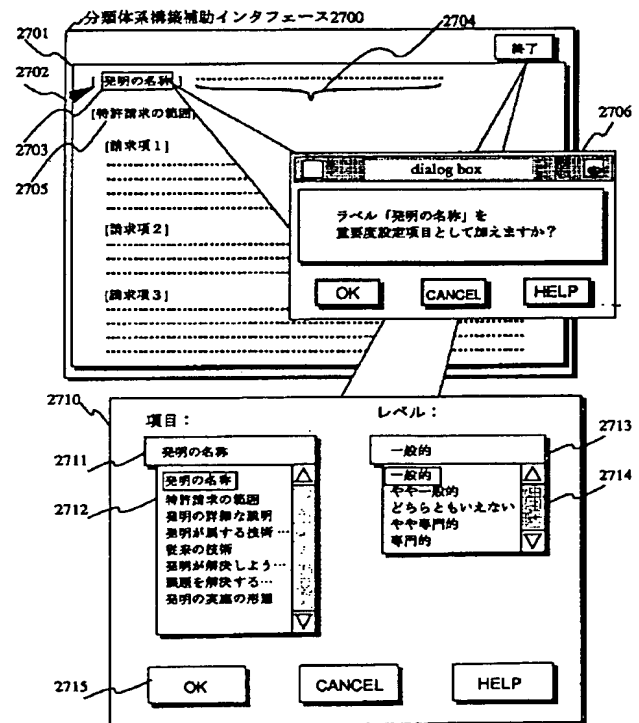
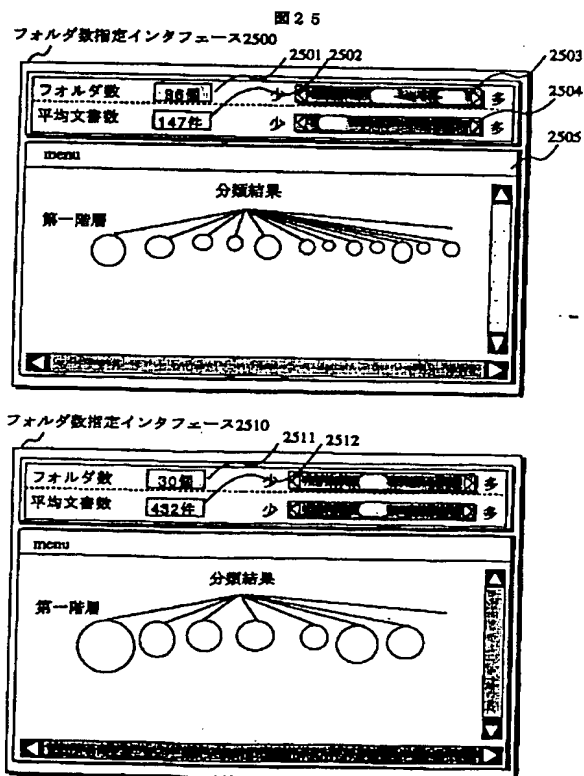
分類指定部 2200



【图 27】

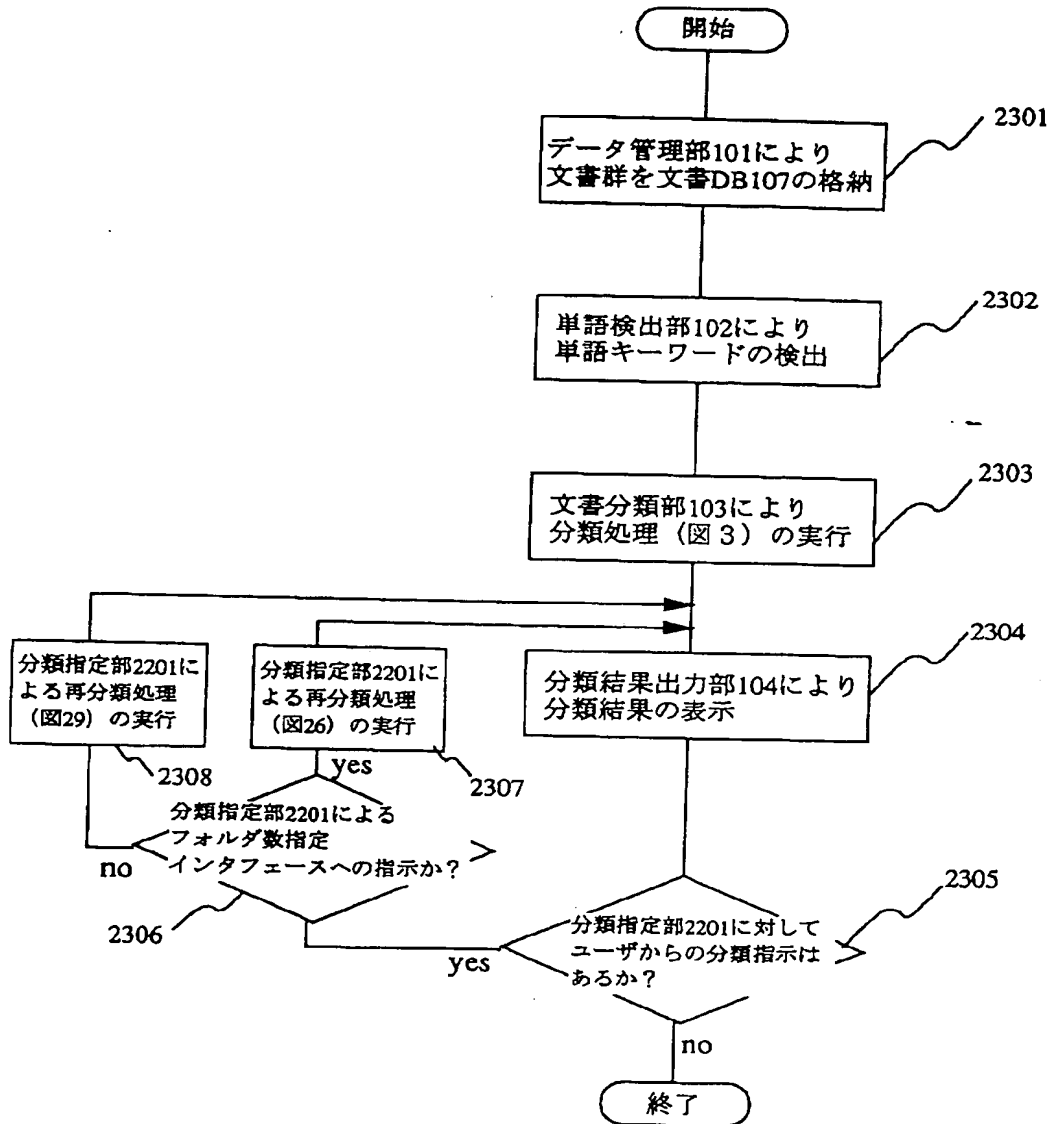
【図25】

图 2-7



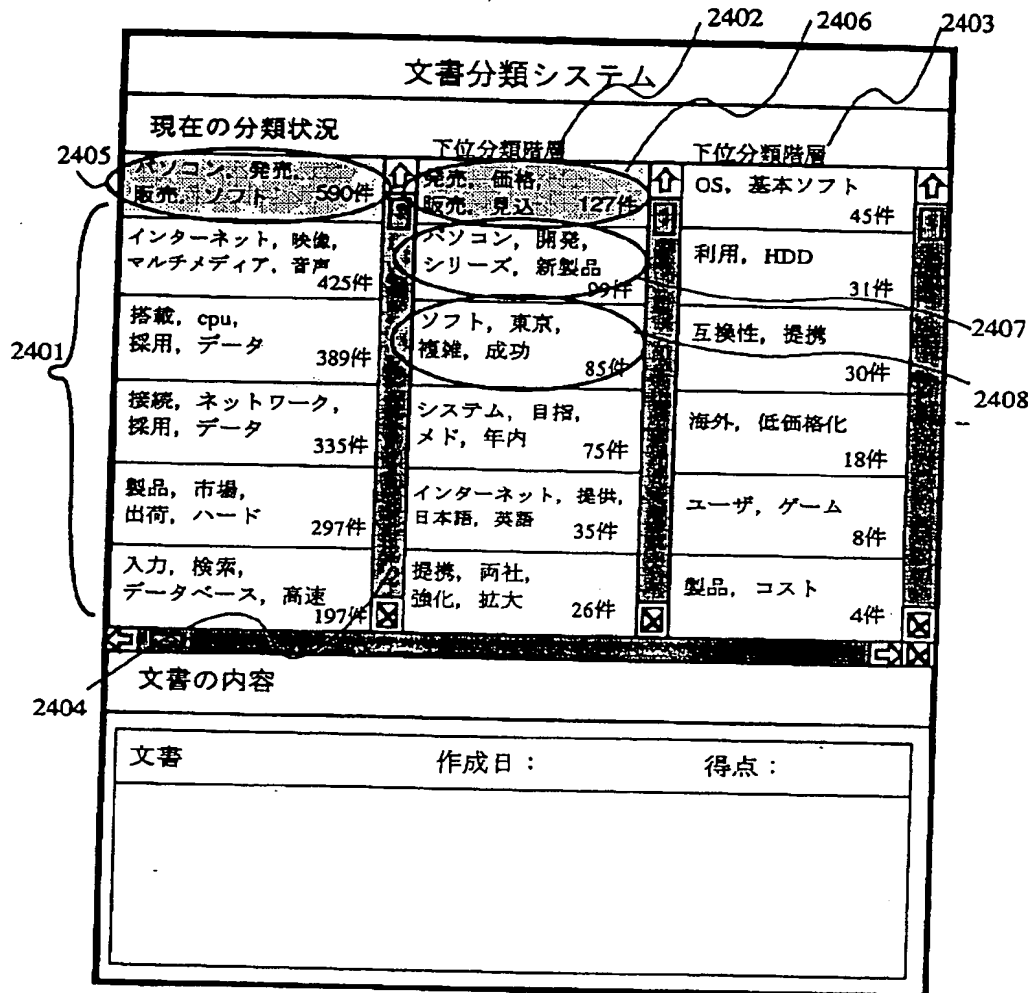
【図23】

図23



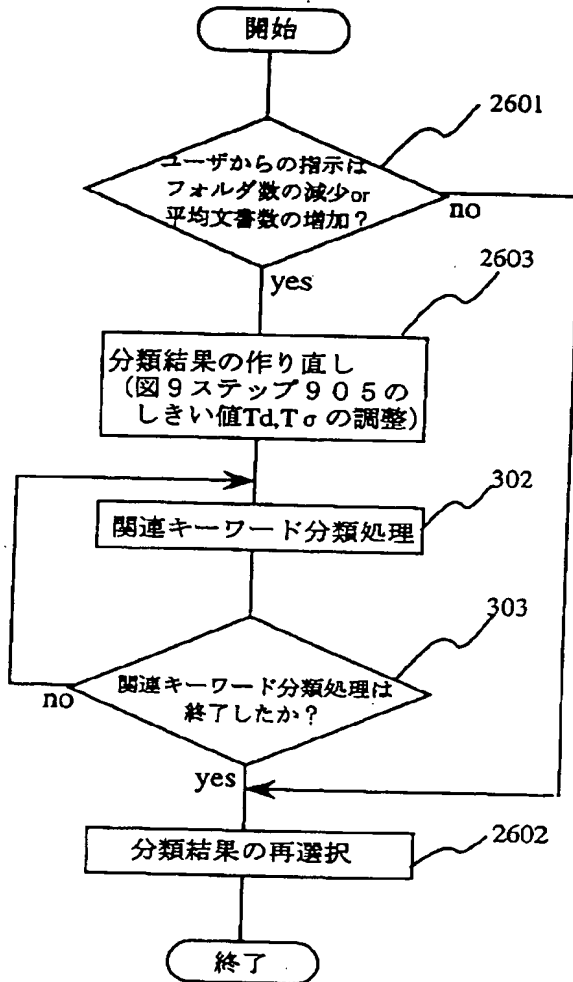
【図24】

図24



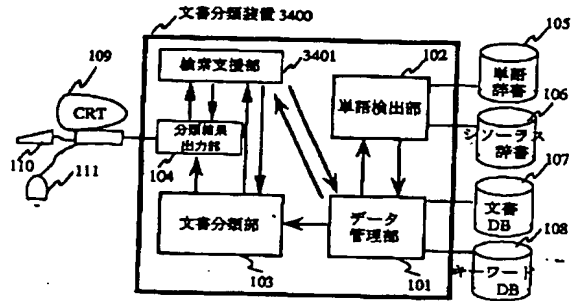
【図26】

図26

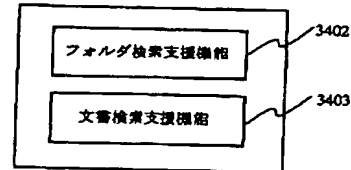


【図34】

図34

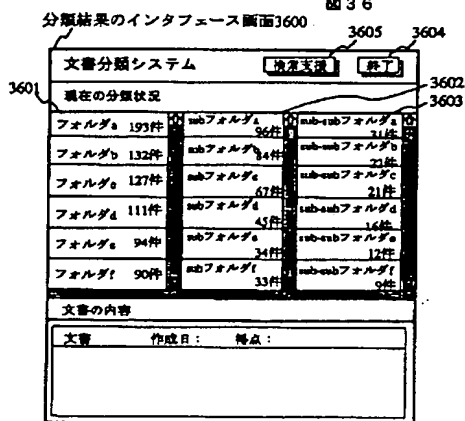


検索支援部 3401

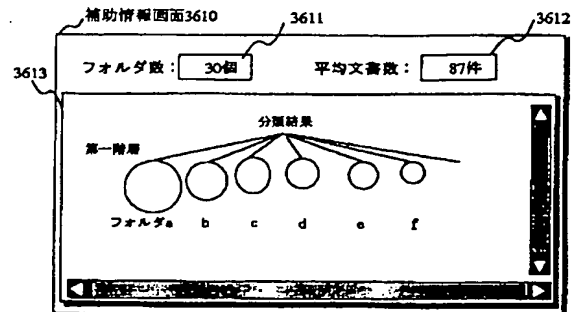


【図36】

図36

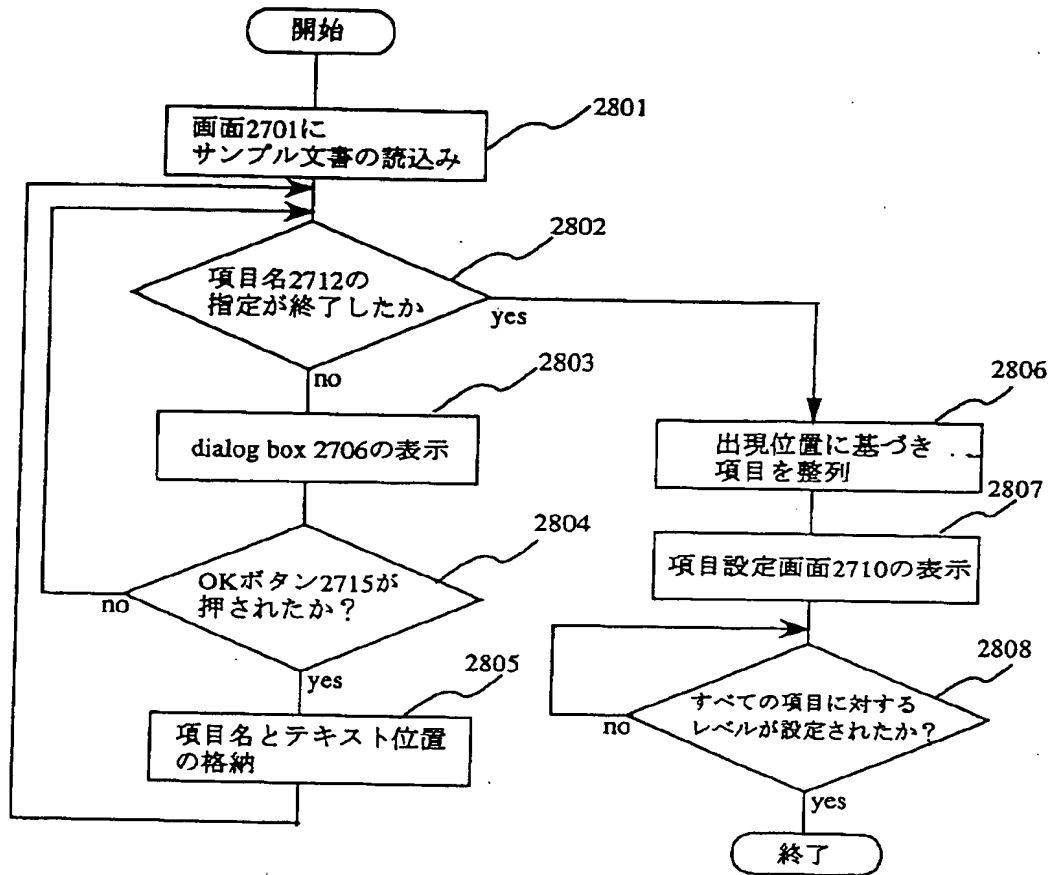


補助情報画面3610



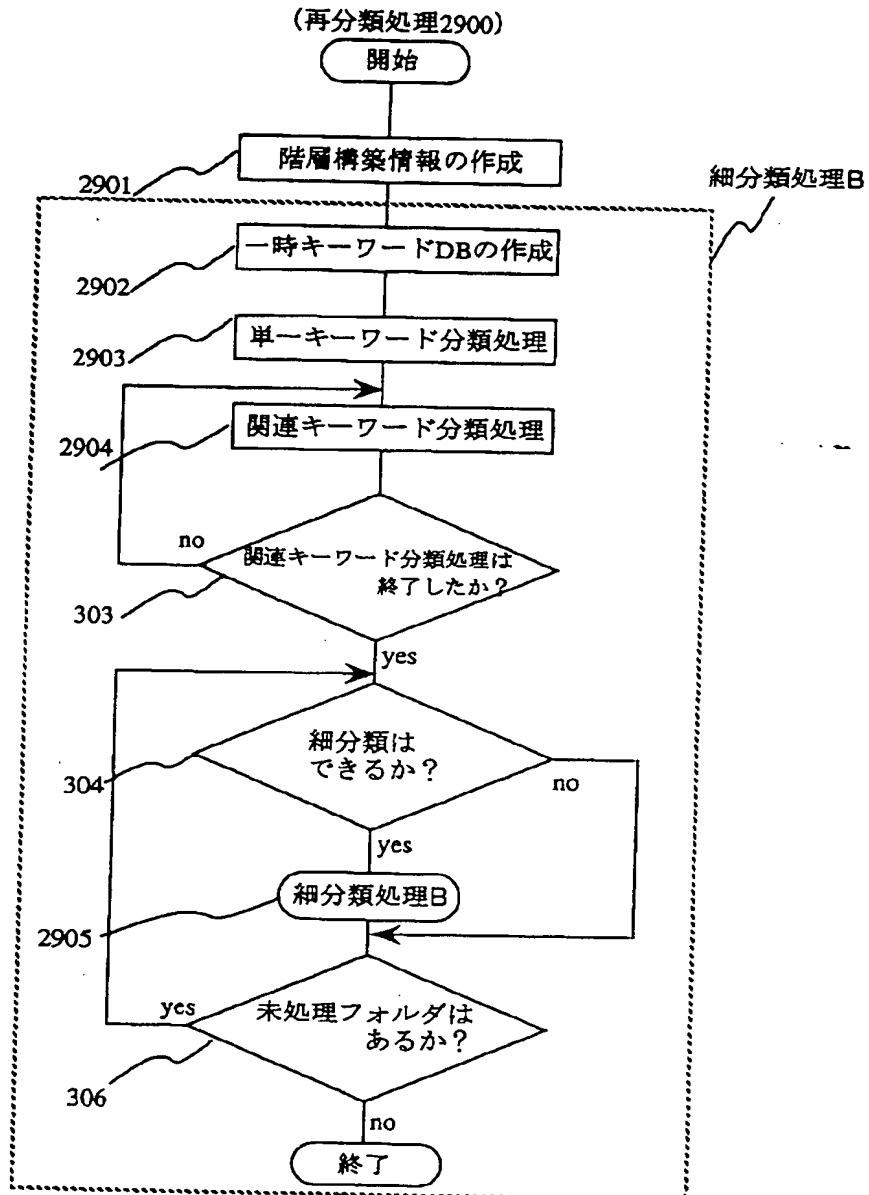
【図28】

図28



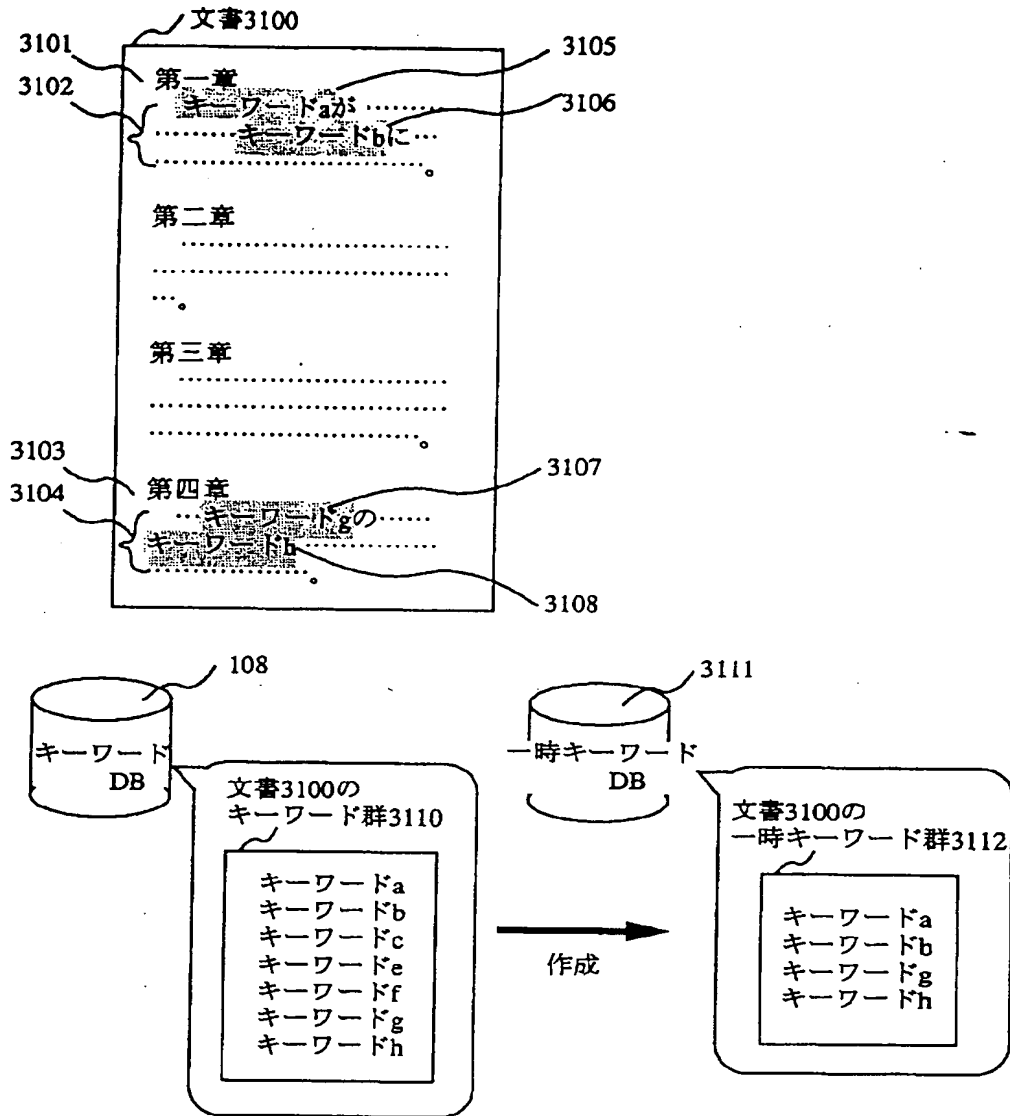
【図29】

図29



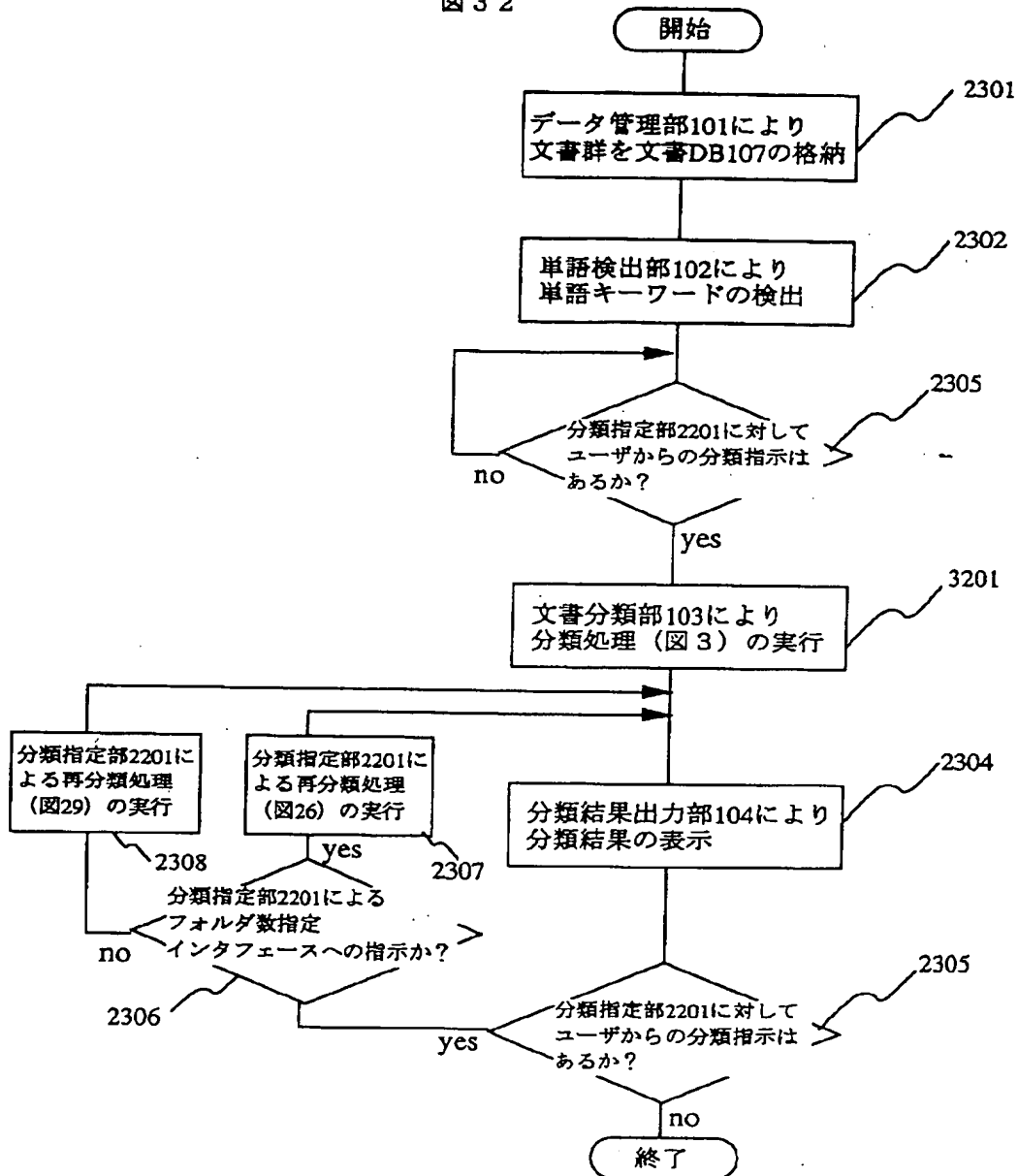
【図31】

図31



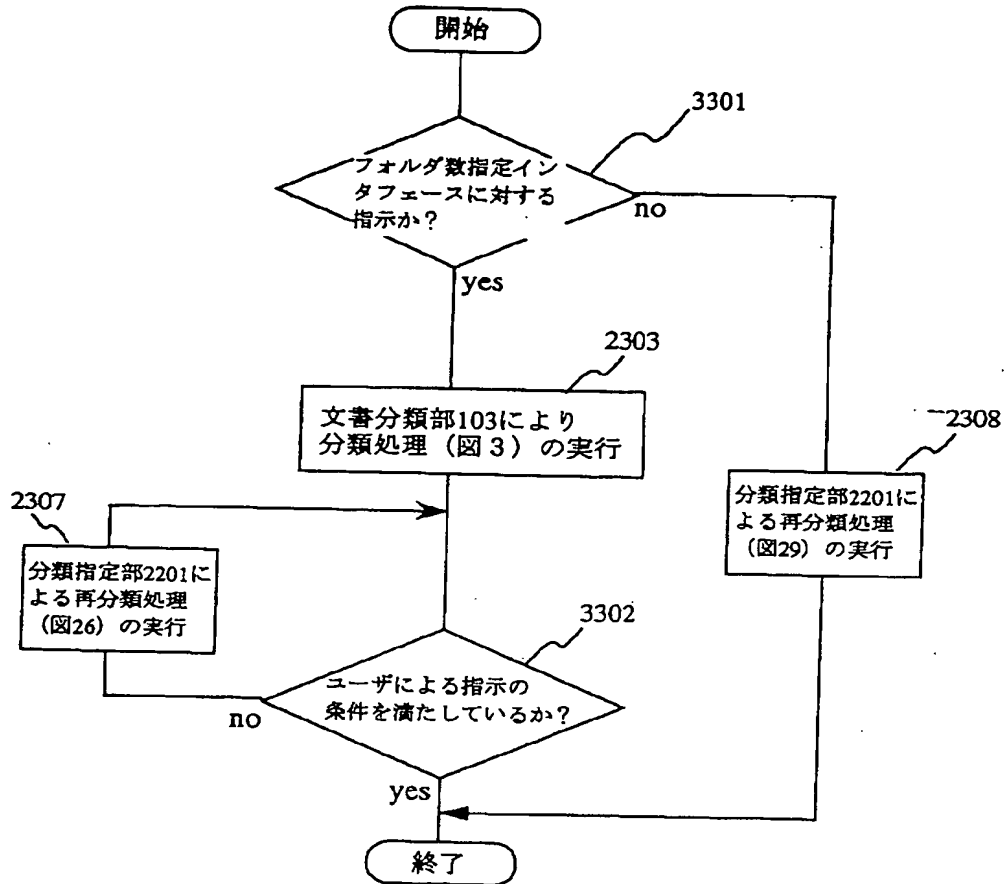
【図32】

図32



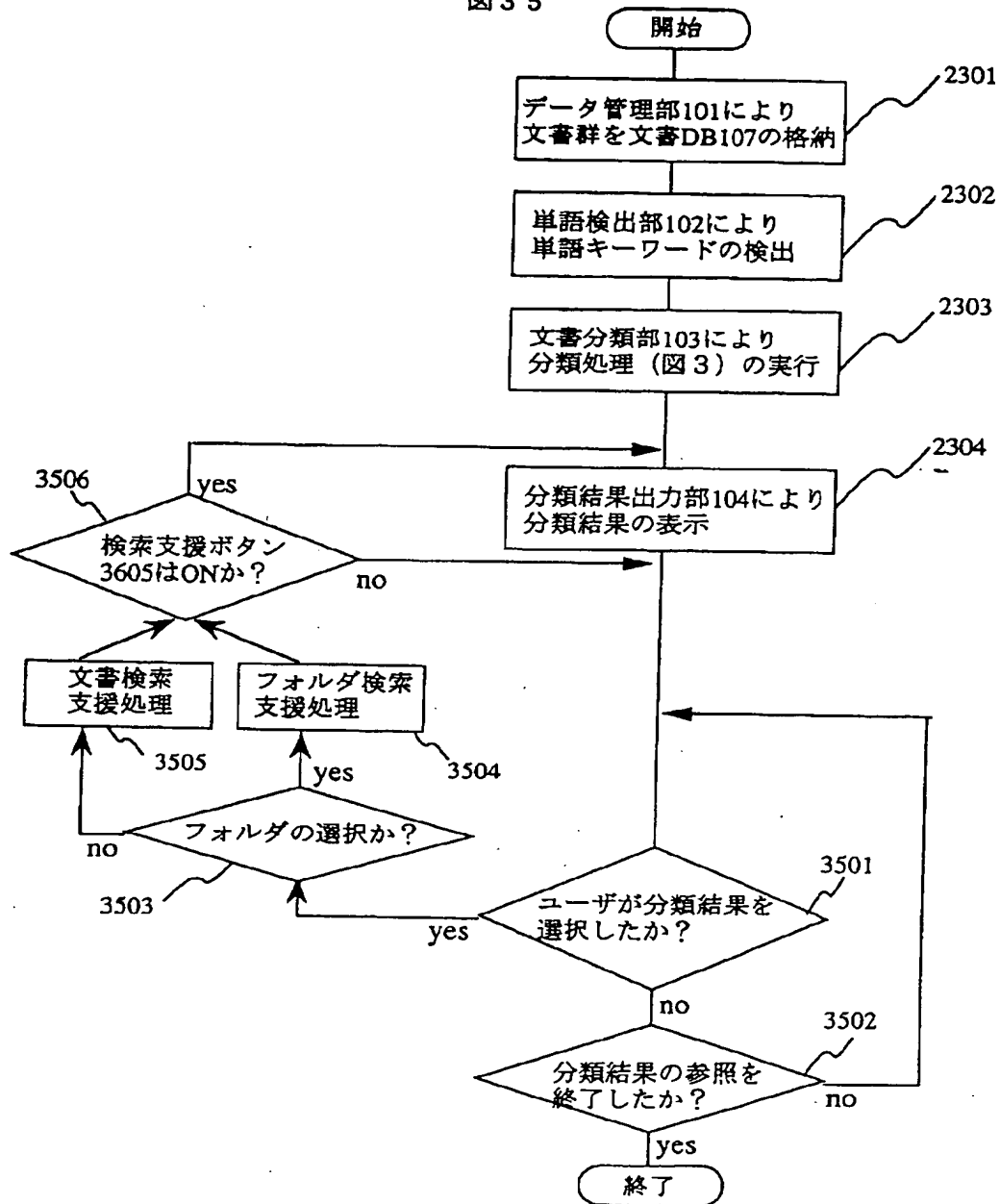
【図33】

図33

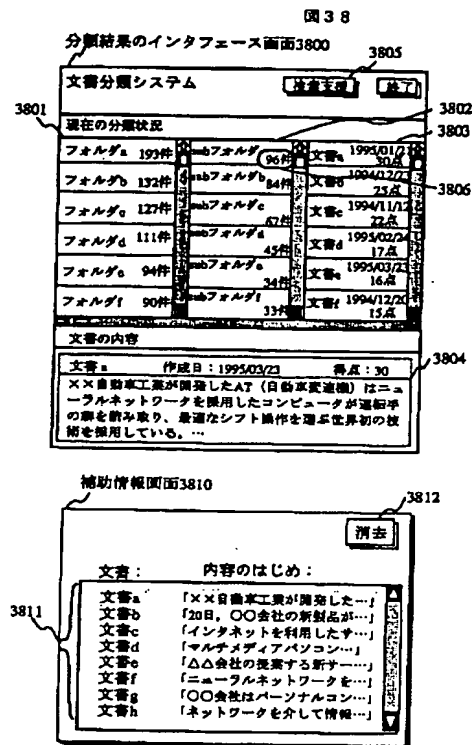


【図35】

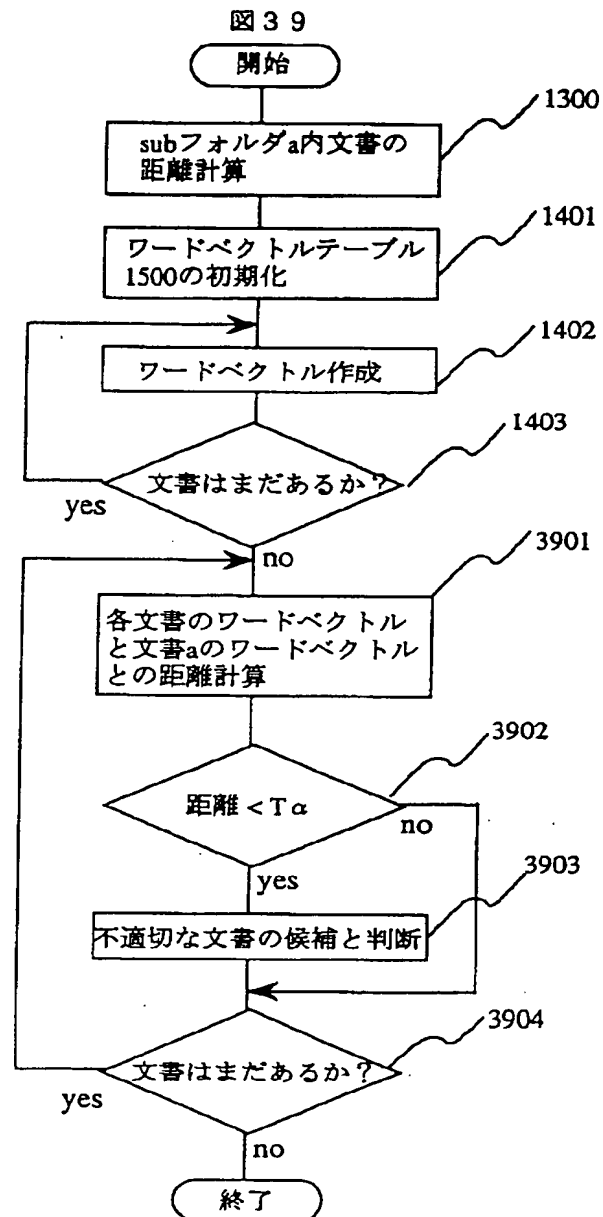
図 3 5



【図38】



【図39】



【図40】

図40

分類結果のインタフェース画面4000

文書分類システム 検索支援 終了

現在の分類状況

フォルダa 193件	subフォルダa 71件	文書b 1994/12/23 25点
フォルダb 132件	subフォルダb 84件	文書c 1994/11/12 22点
フォルダc 127件	subフォルダc 67件	文書c 1995/03/23 16点
フォルダd 111件	subフォルダd 45件	文書g 1995/04/12 11点
フォルダe 94件	subフォルダe 34件	文書h 1995/02/11 11点
フォルダf 90件	subフォルダf 33件	文書i 1995/01/12 9点

4002

4001

文書の内容

文書	作成日	得点

フロントページの続き

(72)発明者 橋本 哲也

 神奈川県川崎市麻生区王禅寺1099番地 株
 式会社日立製作所システム開発研究所内

**An Automatic Document Classification Method Based on
a Semantic Category Frequency Analysis**

by Atsuo Kawai

5

Summary:

Conventional automatic document classification methods can be classified into two groups: (1) a method using statistical information of word notations or kanji notations, and (2) a method using language information and knowledge dependent on a classification system. Although method (1) has high generality due to the independence of its processing from its classification system, it has low precision because it uses only superficial information of word notations, etc. On the other hand, although method (2) has high precision because it uses information dependent on its classification system, it has low generality. This paper proposes a method of applying method (1) to a semantic category system generated independently of its classification system. In this method, first, a system automatically learns a semantic category which is biased and appears in each category field, based on a category example (a

document and its classification system). Then, unclassified documents are classified by using this learning result. In this method, generality can be secured in a sense that even when the classification
5 method is changed, it is sufficient only if a corresponding category example (document for learning) is prepared, and if there is no need to modify the semantic category system used. When its recall ratio is compared with that of a conventional method using
10 only word notations in an experiment where newspaper articles are targeted, the recall ratio has been improved by 10 to 12 %, thereby verifying its effectiveness.

15 1. Introduction

In this paper automatic document classification is studied. There are two document classification methods that can be used with a computer: a method of
20 classifying documents into category fields prescribed in advance, and a method of automatically generating classification system frameworks from given unclassified document groups called a clustering. This paper studies the former method.

25 The method of classifying documents into category

fields prescribed in advance is further roughly divided into two groups: (1) a method using statistical processing (independent of a classification system), and (2) a method using language information and knowledge dependent on a classification system.

In method (1), a system automatically learns words, etc., which are biased and appear in each category field (hereinafter called field-distinguished words), based on the frequency statistics words and kanji notations in a document already classified. Then, the category of the document is determined based on the field-distinguished words in an unclassified document. Since, even when a classification system is changed, it is sufficient only if corresponding classified sample data have been prepared, this method has high generality. However, since only superficial information of word notations, etc., are used in this method, it has low precision.

In method (2), a classification system to be processed is fixed, and classification is performed by using information dependent on the classification system. To be more specific, it includes a method in which classification criteria are incorporated in a system by using generation rules, and a method of

classifying by following hierarchical correspondence tables between keywords, key concepts, and classification fields. Although this method has high precision, it has low generality, since the generation
5 rules and correspondence tables between keywords and classification fields have to be entirely modified manually according to the extension and change of classification system.

Therefore, in this paper, a method utilizing a
10 thesaurus prepared independently of a classification system is proposed in order to secure generality for a classification system and to improve classification precision by using non-superficial information.

In Chapter 2 automatic document classification
15 is studied, and in Chapter 3 the results of experiments using this method are studied, including a comparison with conventional methods.

3. Experiments and Study

20

3.3 Evaluation criteria

A conformity ratio (indicating the size of a classification noise) and a recall ratio (indicating
25 the size of classification omission) are used as

evaluation criteria, as in the case of information retrieval.

$$\text{Conformity ratio} = \frac{\text{RETrel}}{\text{RET}} \cdot 100$$

5 $\text{Recall ratio} = \frac{\text{RETrel}}{\text{rel}} \cdot 100$

where

RET = Total of category candidates outputted by
a system (in all documents)

10 rel = Total of correct categories
 = Total number of all documents to be
processed (Note: This is because one document is
provided with one correct category.)

15 RETrel = Total of correct categories outputted
by a system

20 If a high threshold is set in equation (5)
described in section 2.2.3, a classification field
candidate with high reliability is focussed on, its
conformity ratio increases, and its recall ratio
decreases. Conversely, if a low threshold is set, the
conformity ratio decreases, and the recall ratio
increases. Therefore, in an experiment a threshold is
not fixed and is variously set between 10 and 50, and
the relation between the recall ratio and conformity
25 ratio is examined.

3.4 Evaluation of precision

Figs. 5.1 and 5.2 show the evaluation of the relation between a conformity ratio and a recall ratio in classification systems 1 and 2, respectively. Compared with the conventional case, where only a word dictionary is used, the recall ratio and conformity ratio are better by 12(10)% and 9(10)%, respectively, when both a semantic category dictionary for classification and a word dictionary are used (figures outside paper theses and those inside parentheses are for systems 1 and 2, respectively, and threshold T_h in equation (5) is assumed to be 25 in both systems).

When compared with the case where only a semantic category dictionary is used, higher precision is obtained when both a semantic category dictionary and a word dictionary are used. (The obtained values are at a maximum in systems 1 and 2 when $t_0=0.5$ and $t_0=0.2$, respectively.) The reasons are studied using "JR" and "KDD" for words and "transportation and traffic" and "information and communication" for classification fields. When a word dictionary is used, for example, two fields can be distinguished by using the word dictionary in the form of "JR (transportation and traffic)" and "KDD (information and

communication)". However, since in the semantic category system used in this experiment, both "JR" and "KDD" are only provided with a "company name" semantic category, the two fields cannot be distinguished. In order to distinguish between these two fields using a semantic category system, more detailed semantic categories have to be established. That is, in this example, the resolution of a semantic category system is too broad for the resolution of category fields. In such a case, a word dictionary is effective, and it improves precision.

In this case, the precision can also be improved by classifying a thesaurus (semantic category system) in more detail. When this customization is compared with the manual composition of a field-dependent thesaurus, with one knowledge expression dependent on the classification system described in method (2) in Chapter 1, there is no difference between them in individual jobs. An example of this is in the setting of the "communication" and "railway transportation" classification fields shown in Fig. 6 and the determination of company names belonging to classification fields.

However, the method studied in this paper is different from method (2) described in Chapter 1, in

the point that an objective score weighting to individual semantic categories due to learning using frequency information is available after the detailed classification of the semantic category system. For example and as shown in Fig. 6, it is assumed that the business field of corporations is classified as a semantic category. In this case, "air transportation" often appears in articles on the transportation and communication field, and it also appears in articles on the diplomacy field. In this way, a "classification field" corresponds to fields in which it appears, one to n , and the degree of its individual relation varies with the classification field to which it belongs. According to the method studied in this paper, this degree of relation (score) can be statistically calculated using a document for learning. On the other hand, in method (2) described in Chapter 1, only semantic categories and words that appear frequently are registered for each field. In order to take into consideration the degree of relation between a classification field and a semantic category or word, a manual intuitive degree of relation must be used.

In method (2) described in Chapter 1, when a classification system is partly modified after customization, corresponding modification is required.

However, in this method, since the score weighting to a semantic category is modified by new learning, there is no need to modify the thesaurus. As an example of this, the classification fields shown in the upper column of Fig. 6 are modified to those shown in the lower column of Fig. 6. In this case, the relation between a classification field newly set and the existing semantic category (including the semantic category shown in Fig. 6) has to be newly established. In method (2) described in Chapter 1, this setting has to be performed manually. However, in the method described in this paper, this relation can be automatically established by using a document for learning based on a new classification system.

There are also documents in which a correct category cannot be obtained, even if both a word dictionary and a semantic category dictionary are used. Out of these documents, three factors, from the viewpoint of word notation and semantic category, are focussed on in this paper, as given below.

(1) There is a document in which no correct category can be obtained because words contained in a document cannot be grouped at the semantic category level, although they can be grouped in a certain point of view. For example, as words which appear in a

document, there are "an unearthed article" ("article"), "an excavation" ("gathering") and "a cultural asset" ("asset"). These words can be grouped in an archaeology classification (field=culture).

5 However, at the semantic category level the words can only be supplied with semantic categories, "article", "excavation" and "asset", and it is difficult to extract a relation, field=culture (archaeology).

(2) Errors result due to processing words with a
10 polysemy (a plurality of semantic categories), while their semantic categories are not narrowed.

(3) Errors result due to adding the score of category candidates, taking into consideration the individual word notation and semantic category of words composing
15 an idiom, etc. (for example, grass root). For example, grass-root personal computer communication: agriculture and forestry and fishery (incorrect category), information communication (correct category).

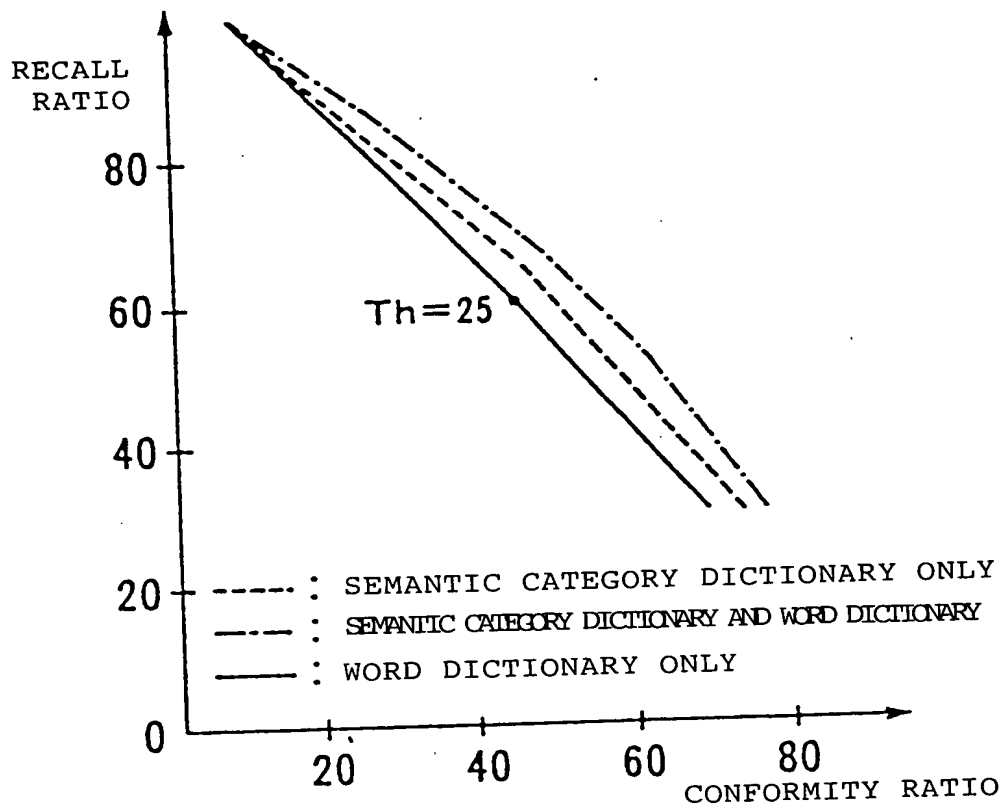


Fig. 5-1 Evaluation of reference dictionaries
(classification system 1) (Recall
Precision Graph).

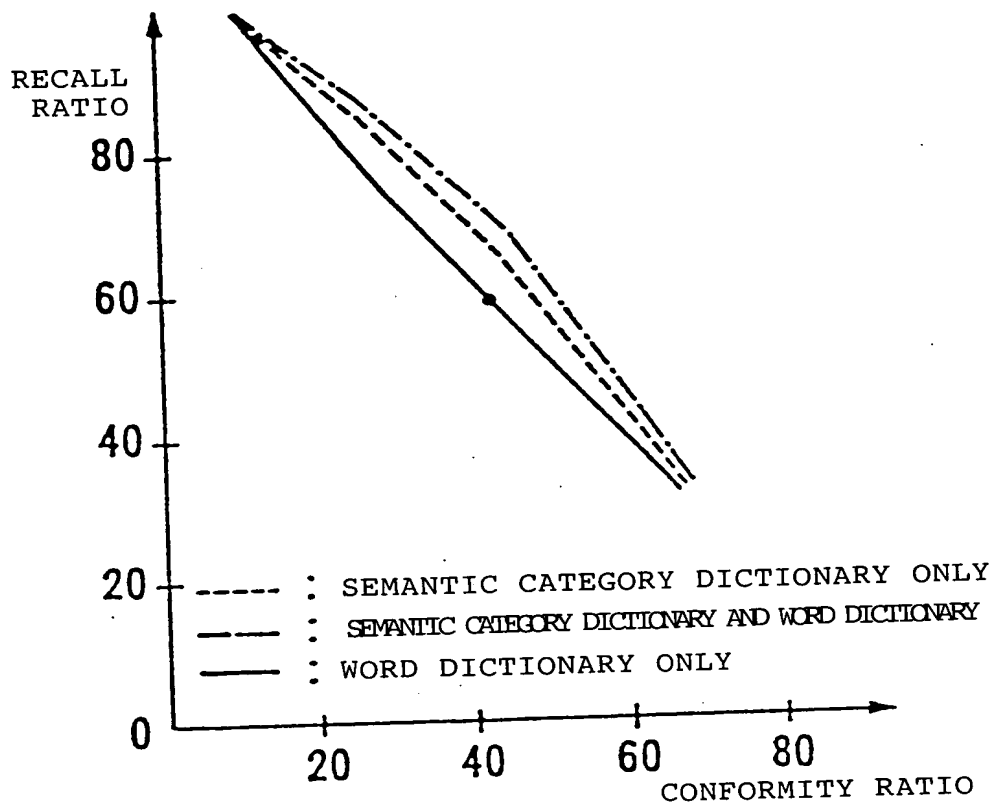


Fig. 5-2 Evaluation of reference dictionaries (classification system 2). (Recall Precision Graph).

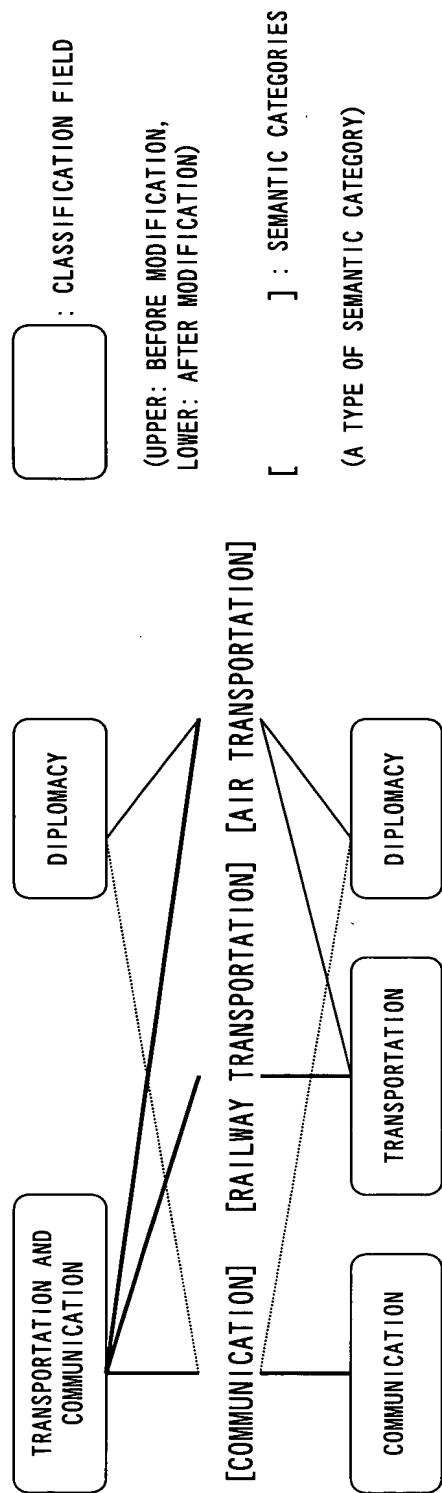


FIG. 6 Relation between classification fields and semantic categories.